University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 142

The Cross-Validated Adaptive Epsilon-Net Estimator

Mark J. van der Laan* Sandrine Dudoit†
Aad W. van der Vaart‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

‡Dept. of Mathematics, Vrije Universitat, Amsterdam

http://biostats.bepress.com/ucbbiostat/paper142

The Cross-Validated Adaptive Epsilon-Net Estimator

Mark J. van der Laan, Sandrine Dudoit, and Aad W. van der Vaart

Abstract

Suppose that we observe a sample of independent and identically distributed realizations of a random variable. Assume that the parameter of interest can be defined as the minimizer, over a suitably defined parameter space, of the expectation (with respect to the distribution of the random variable) of a particular (loss) function of a candidate parameter value and the random variable. Examples of commonly used loss functions are the squared error loss function in regression and the negative log-density loss function in density estimation. Minimizing the empirical risk (i.e., the empirical mean of the loss function) over the entire parameter space typically results in ill-defined or too variable estimators of the parameter of interest (i.e., the risk minimizer for the true data generating distribution). In this article, we propose a cross-validated epsilon-net estimation methodology that covers a broad class of estimation problems, including multivariate outcome prediction and multivariate density estimation. An epsilon-net sieve of a subspace of the parameter space is defined as a collection of finite sets of points, the epsilon-nets indexed by epsilon, which approximate the subspace up till a resolution of epsilon. Given a collection of subspaces of the parameter space, one constructs an epsilon-net sieve for each of the subspaces. For each choice of subspace and each value of the resolution epsilon, one defines a candidate estimator as the minimizer of the empirical risk over the corresponding epsilon-net. The cross-validated epsilon-net estimator is then defined as the candidate estimator corresponding to the choice of subspace and epsilon-value minimizing the cross-validated empirical risk. We derive a finite sample inequality which proves that the proposed estimator achieves the adaptive optimal minimax rate of convergence, where the adaptivity is achieved by considering epsilon-net sieves for various subspaces. We also address the implementation of the cross-validated epsilon-net estimation procedure. In the context of a linear regression model, we present results of a preliminary simulation

study comparing the cross-validated epsilon-net estimator to the cross-validated L1-penalized least squares estimator (LASSO) and the least angle regression estimator (LARS). Finally, we discuss generalizations of the proposed estimation methodology to censored data structures.

Contents

1 Introduction

1.1 Loss-based estimation

Parameters and loss functions. Let $O_1, \ldots, O_n$ be n independent and identically distributed (i.i.d.) observations from a data generating distribution P_0, known to be an element of a statistical model $\mathcal{M}$. Let $\Psi : \mathcal{M} \to D(\mathcal{S})$ denote a *parameter*, i.e., a mapping from the model $\mathcal{M}$ into a space $D(\mathcal{S})$ of real-valued functions from a Euclidean set $\mathcal{S} \subseteq \mathbb{R}^d$. We denote the *parameter space* corresponding to this parameter with $\mathbf{\Psi} \equiv \{\Psi(P) : P \in \mathcal{M}\} \subseteq D(\mathcal{S})$. Each *parameter value* $\psi \in \mathbf{\Psi}$ is therefore a function, $\psi : \mathcal{S} \to \mathbb{R}$, from a certain d-dimensional Euclidean set $\mathcal{S} \subseteq \mathbb{R}^d$ into the real line. In particular, let $\psi_0 \equiv \Psi(P_0)$ be the *true parameter value* (i.e., a function) corresponding to the data generating distribution P_0. Note that the use of upper case Ψ and lower case ψ allows us to distinguish between the mapping $\Psi : \mathcal{M} \to \mathcal{D}(\mathcal{S})$ and actual realizations ψ of this mapping which are themselves functions from $\mathcal{S}$ into the real line. We assume that the space $D(\mathcal{S})$ is endowed with a *dissimilarity function*, $d : D(\mathcal{S}) \times D(\mathcal{S}) \to \mathbb{R}$, defining the dissimilarity $d(\psi_1, \psi_2)$ between two elements ψ_1 and ψ_2 of $\mathbf{\Psi}$.

Let $L : (O, \psi) \to L(O, \psi) \in \mathbb{R}$ be a *loss function* which maps a candidate parameter value $\psi \in \mathbf{\Psi}$ and observation O into a real number, and whose expected value (i.e., risk) is minimized at the parameter value $\psi_0 = \Psi(P_0)$ corresponding to the data generating distribution P_0. That is,

$$\psi_0 = \text{argmin}_{\psi \in \mathbf{\Psi}} E_{P_0} L(O, \psi) = \text{argmin}_{\psi \in \mathbf{\Psi}} \int L(o, \psi) dP_0(o). \quad (1)$$

We adopt terminology from the prediction literature and define the *risk* of a candidate $\psi \in \mathbf{\Psi}$ as the expected loss $E_{P_0} L(O, \psi)$. We refer to the difference between the risk at ψ and the minimal risk at ψ_0 as the *risk difference at* ψ

$$d_0(\psi, \psi_0) \equiv E_{P_0}[L(O, \psi) - L(O, \psi_0)] = \int (L(o, \psi) - L(o, \psi_0)) dP_0(o). \quad (2)$$

Estimators. Let P_n denote the *empirical distribution* of $O_1, \ldots, O_n$, where P_n places probability $1/n$ on each realization O_i, $i = 1, \ldots, n$. Our goal is to use the sample to estimate the parameter $\psi_0 = \Psi(P_0)$ of the unknown data generating distribution P_0. An *estimator* $\hat{\Psi}$ is a mapping from empirical distributions to the parameter space $\mathbf{\Psi}$ and a realization of this mapping will be denoted by $\hat{\psi} \equiv \hat{\Psi}(P_n)$. Note that estimators $\hat{\Psi}$ are viewed as algorithms

one can apply to *any* empirical distribution, not as the actual realizations $\hat{\psi}$ at the observed P_n. We refer to

$$\int L(o, \hat{\Psi}(P_n))dP_0(o)$$

as the *conditional risk* (given the empirical distribution P_n) of the estimator $\hat{\psi} = \hat{\Psi}(P_n)$. The expectation of this conditional risk is referred to as the *marginal risk*. Finally,

$$\int L(o, \hat{\Psi}(P_n))dP_n(o)$$

is called the *empirical risk* estimator (or resubstitution risk estimator) and can be viewed as an estimator of both conditional risk as well as marginal risk.

Given a guessed collection of user-supplied parameter subspaces $\mathbf{\Psi}_s \subseteq \mathbf{\Psi}$, with $\mathbf{\Psi} = \mathbf{\Psi}_s$ for some s, our goal is to construct an estimator $\Psi(P_n)$ of ψ_0 whose marginal risk minus the minimal risk, i.e., whose expected risk difference $Ed_0(\hat{\Psi}(P_n), \psi_0)$, converges to zero at a rate which is at worst equal to the minimax rate implied by the size of the smallest of the parameter subspaces which contains the true ψ_0. Such an estimator is called *adaptive* (e.g., Barron et al. (1999), Birgé and Massart (1997), Yang and Barron (1999)). Our proposed estimation framework covers, in particular, multivariate regression and multivariate density estimation. Other important problems, including loss-based estimation with censored data, are discussed in related articles (Dudoit and van der Laan, 2003; Dudoit et al., 2004; Keleş et al., 2003; Molinaro et al., 2004; Molinaro and van der Laan, 2004; Sinisi and van der Laan, 2004; van der Laan et al., 2004; van der Laan and Dudoit, 2003).

Univariate outcome regression. Let $O = (W, Y) \sim P_0$, where Y is a scalar outcome and W is a vector of covariates with cumulative distribution function (c.d.f.) F_0. The parameter of interest is the conditional expected value, $\psi_0(W) \equiv E_{P_0}[Y \mid W]$, of the outcome Y given covariates W. We can use as loss function the *quadratic loss function*,

$$L(O, \psi) \equiv (Y - \psi(W))^2, \tag{3}$$

also known as the *squared error loss function* or L^2 *loss function*. Note that the risk difference $d_0(\psi, \psi_0)$ equals the expected value of the squared

difference between the candidate ψ and the truth ψ_0, that is,

$$d_0(\psi, \psi_0) = E_{P_0}\left[(\psi(W) - \psi_0(W))^2\right] = \int (\psi(w) - \psi_0(w))^2 dF_0(w). \quad (4)$$

Multivariate outcome regression. Similarly, for multivariate outcome regression, let $O = (W, Y) \sim P_0$, where $Y = (Y(l) : l = 1, \ldots, L)$ is a random outcome L-vector and W a vector of covariates with c.d.f. F_0. The parameter of interest is $\psi_0(W) \equiv E_{P_0}[Y \mid W] = (E_{P_0}[Y(l) \mid W] : l = 1, \ldots, L)$, the conditional expected value of the outcome vector Y given covariates W. Define the loss function as the *weighted quadratic loss function*,

$$L(O, \psi) \equiv (Y - \psi(W))^\top \eta(W)(Y - \psi(W)), \quad (5)$$

where $\eta(\cdot)$ is a symmetric $L \times L$-matrix function of W, and note that, for any choice of $\eta(\cdot)$, the risk is minimized by the parameter value ψ_0, that is,

$$\psi_0 = \text{argmin}_{\psi \in \mathbf{\Psi}} E_{P_0} L(O, \psi) = \text{argmin}_{\psi \in \mathbf{\Psi}} \int L(o, \psi) dP_0(o).$$

Here, $\eta(W)$ could denote an approximation of an unknown matrix, such as the inverse of the conditional covariance matrix $\Sigma(W)$ of the outcome Y given covariates W,

$$\Sigma(W) \equiv E_{P_0}\left[(Y - E_{P_0}[Y \mid W])\,(Y - E_{P_0}[Y \mid W])^\top \mid W\right].$$

Density estimation. Finally, if $O \sim f_0 \equiv \frac{dP_0}{d\mu}$, where μ is a dominating measure of the data generating distribution P_0, and the density function $\psi_0(O) \equiv f_0(O)$ is the parameter of interest, then we can define the loss function as the *negative log-density loss function*,

$$L(O, \psi) \equiv -\log(\psi(O)). \quad (6)$$

Note that the minimum risk $E_{P_0} L(O, \psi_0)$ is the *entropy* of the distribution P_0 and the risk difference $d_0(\psi, \psi_0)$ equals the *Kullback-Leibler divergence* between the candidate ψ and the true density ψ_0.

1.2 ϵ-net sieves

The minimum empirical risk estimator (e.g., least squares estimator or maximum likelihood estimator in the examples above)

$$\text{argmin}_{\psi \in \mathbf{\Psi}} \int L(o, \psi) dP_n(o)$$

might not be well-defined and/or might suffer from the curse of dimensionality. A general approach advocated throughout the literature for dealing with this problem is too construct a sequence of subspaces approximating the whole parameter space $\mathbf{\Psi}$, a so-called *sieve*, and then select the subspace whose corresponding minimum empirical risk estimator minimizes an appropriately penalized empirical risk or cross-validated empirical risk.

A general loss function-based approach for model selection and estimation, described in Barron et al. (1999), uses sieve theory to define penalized empirical risk criteria. In particular, Barron (1989) and Barron (1991) develop this theory in the context of artificial neural networks. Connections with cross-validation methods are discussed in Birgé and Massart (1997). Barron et al. (1999) and Birgé and Massart (1997) have studied thoroughly the penalty functions to be used in adaptive estimation on sieves. They use powerful Talagrand concentration and deviation inequalities for empirical processes (Ledoux, 1996; Massart, 1998; Talagrand, 1996a,b) to obtain so-called oracle inequalities for the theoretical risk of their estimators. The method of oracle inequalities was also used to prove optimality properties of non-parametric estimators in Johnstone (1998). The Birgé-Massart penalties are based on the dimension of the classes of functions. This approach has been shown to perform well for sieves that frequently occur in non-parametric univariate regression and non-parametric univariate density estimation problems (e.g., nested families of Sobolev ellipsoids).

In this article, we focus on a particular type of sieve, namely *ϵ-nets* $\{\mathbf{\Psi}_{s,\epsilon} : \epsilon\}$ of guessed subspaces $\mathbf{\Psi}_s$ of the complete parameter space $\mathbf{\Psi}$, and rely on *cross-validated risk estimation* to define the subspace and resolution selection criteria, i.e., to select the pair (s, ϵ). Let $\hat{\Psi}_{s,\epsilon}(P_n)$ be the minimum empirical risk estimator for the (s, ϵ)-specific ϵ-net $\mathbf{\Psi}_{s,\epsilon}$. Our proposed *cross-validated ϵ-net estimator* can be denoted as $\hat{\Psi}(P_n) = \hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n)$, where $(s(P_n), \epsilon(P_n))$ is the minimizer of the cross-validated empirical risk over all candidate estimators $\hat{\Psi}_{s,\epsilon}(P_n)$. We prove a general finite sample inequality for the expected risk difference $Ed_0(\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_{n(1-p)}), \psi_0)$, i.e., the *marginal risk* of the data-adaptively selected estimator $\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n)$, applied to a subsample of size $n(1-p)$, minus the *minimal risk*, where p denotes the proportion of observations constituting the validation sample in the employed cross-validation scheme (Theorems 1 and 3). This finite sample inequality teaches us that the proposed estimator achieves at worst the *minimax rate* implied by the size of the parameter space $\mathbf{\Psi}$. In addition, since

the cross-validated ϵ-net estimator $\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n)$ chooses data-adaptively (using cross-validation) both the subspace $\mathbf{\Psi}_s$ and resolution ϵ, the finite sample inequality demonstrates that this estimator is *adaptive*. Our previous results on cross-validation selection show that the cross-validated choice of the (s, ϵ)-pair is typically asymptotically equivalent with an *oracle* procedure making the optimal P_0-dependent choice for (s, ϵ) for the given dataset (van der Laan and Dudoit, 2003). Thus, cross-validation is a very adaptive procedure and thereby might be preferable to the use of universal penalty terms that are independent of the true distribution P_0. Regarding our choice of sieve, and as discussed below, the *sparsity* of ϵ-nets makes ϵ-net sieves particularly effective.

Le Cam has used ϵ-nets in the context of parametric models to construct efficient estimators under minimal conditions; this approach is often referred to as Le Cam's discretization device (Le Cam, 1986; Le Cam and Yang, 1990). Our argument in favor of the use of ϵ-net sieves is that, by definition, an ϵ-net equals the smallest (in size, and thereby sparsest) ϵ-approximation of the complete parameter space. In contrast, commonly used sieves might yield dense approximations in certain areas of the parameter space, but might result in ineffective approximations in other parts of the space. Recently, Donoho (2003) has argued a theoretical geometrical advantage of ϵ-nets in relation to other choices of sieves in the context of univariate non-parametric regression. This advantage of ϵ-nets is connected with the sparsity concept developed in Donoho and Johnstone (1994).

We stress that existing approaches in the statistics and machine learning literatures for non-parametric multivariate regression and conditional density estimation do not rely on ϵ-net sieves (Breiman et al., 1984; Hastie et al., 2001; Ripley, 1996). For example, commonly used sieves correspond with constraints on the norm of the vector of coefficients. Furthermore, these approaches do not aim to minimize the empirical mean of the loss function (e.g., sum of squared residual errors) over specified subspaces of the complete parameter space. Instead, current algorithms rely on forward/backward-like local optimization steps. The present article provides a road map for developing minimax adaptive estimators in a large class of estimation problems based on ϵ-net sieves and cross-validation.

1.3 Overview

Section 2 formally defines the proposed cross-validated adaptive ϵ-net estimation approach. It also provides a practical methodology for the construction of an ϵ-net sieve and a straightforward algorithm for minimizing the empirical risk over an ϵ-net. Section 3 establishes the theoretical foundations of the proposed estimation methodology. We prove two main theorems which provide finite sample bounds for the expected risk difference, $Ed_0(\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_{n(1-p)}), \psi_0)$, between the proposed cross-validated ϵ-net estimator $\hat{\Psi}(P_n) = \hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n)$ and the parameter value $\Psi(P_0) = \psi_0$. The first theorem (Theorem 1) concerns loss functions whose risk at ψ_0 can be estimated at a rate quadratic in the rate at which ψ_0 itself can be estimated, while the second theorem (Theorem 3) concerns general loss functions. We demonstrate that the finite sample inequality of Theorem 1 implies optimal rates of convergence and adaptivity in various general examples. We refer to Yang and Barron (1999) for the formal result showing that our finite sample inequality indeed, in general, implies the minimax rate of convergence. Corollaries of the finite sample result are provided in Section 4 for multivariate regression and density estimation. In particular, the finite sample inequality is illustrated for the cases that: (i) the ϵ-net sieves approximate well-known smoothness classes for multivariate real-valued functions indexed by s (one of them being the complete parameter space); and (ii) the ϵ-net sieves approximate linear regression models with maximally s variables. In the context of linear regression, Section 5 compares the practical performance of the cross-validated ϵ-net estimator with L^1-penalized least squares linear regression (LASSO, Tibshirani (1996)), where the penalty is chosen with cross-validation, and least angle linear regression (LARS, Efron et al. (2004)). Finally, Section 6 refers to extensions of our theorems to loss functions which depend on a nuisance parameter (van der Laan and Dudoit, 2003), thereby covering a whole range of new and important estimation problems. In particular, a link to censored data estimation theory, as presented in van der Laan and Robins (2002), provides a class of loss functions for estimation based on censored data.

2 The cross-validated adaptive ϵ-net estimator

2.1 Outline of the cross-validated ϵ-net estimation procedure

Our proposed cross-validated adaptive ϵ-net estimation approach consists of the following three main steps. We elaborate on each of these steps below.

1. Define a collection of subspaces, $\mathbf{\Psi}_s \subseteq \mathbf{\Psi}$, and corresponding ϵ-net sieves, $\{\mathbf{\Psi}_{s,\epsilon} : \epsilon\}$.

2. For each choice of subspace $\mathbf{\Psi}_s$ and ϵ-net resolution ϵ, construct candidate estimators as the empirical risk minimizers over the ϵ-nets $\mathbf{\Psi}_{s,\epsilon}$.

3. Select the subspace $\mathbf{\Psi}_s$ and ϵ-net resolution ϵ by cross-validation.

2.1.1 Definition of subspaces and ϵ-net sieves

Let $\mathbf{\Psi}_s \subseteq \mathbf{\Psi}$, $s = 1, \ldots, K_1(n)$, be a collection of subspaces, where one or more of these subspaces equals the complete parameter space $\mathbf{\Psi}$.

Definition 1 (ϵ-nets and ϵ-net sieves) *For any subspace $\mathbf{\Psi}_s$ of $\mathbf{\Psi}$ and resolution $\epsilon > 0$, an* ϵ-net *$\mathbf{\Psi}_{s,\epsilon}$ of $\mathbf{\Psi}_s$ is defined as a finite subset of $\mathbf{\Psi}_s$,*

$$\mathbf{\Psi}_{s,\epsilon} \equiv \{\psi_1^{s,\epsilon}, \ldots, \psi_{N_s(\epsilon)}^{s,\epsilon}\} \subseteq \mathbf{\Psi}_s, \tag{7}$$

so that the union $\cup_{j=1}^{N_s(\epsilon)} B(\psi_j^{s,\epsilon}, \epsilon)$, of all spheres $B(\psi_j^{s,\epsilon}, \epsilon) \equiv \{\psi \in \mathbf{\Psi}_s : d(\psi, \psi_j^{s,\epsilon}) \leq \epsilon\}$ centered at $\psi_j^{s,\epsilon}$ with radius ϵ, covers the parameter space $\mathbf{\Psi}_s$. A collection $\{\mathbf{\Psi}_{s,\epsilon} : \epsilon\}$ of such ϵ-nets is an ϵ-net sieve *of $\mathbf{\Psi}_s$, $s = 1, \ldots, K_1(n)$.*

Definition 2 (Covering number) *The* covering number *of the subspace $\mathbf{\Psi}_s$ of $\mathbf{\Psi}$ is defined as the minimal number $N(\epsilon, \mathbf{\Psi}_s, d)$ of spheres needed to cover $\mathbf{\Psi}_s$ at a resolution of $\epsilon > 0$. This function of ϵ should be viewed as a measure of the* size *of the parameter space $\mathbf{\Psi}_s$ (van der Vaart and Wellner, 1996).*

Note that we can have $\mathbf{\Psi}_s = \mathbf{\Psi}$ for several s, where different choices of subspace $\mathbf{\Psi}_s$ yield ϵ-nets $\mathbf{\Psi}_{s,\epsilon}$ corresponding to different parameterizations of

$\mathbf{\Psi}$. As an example, $\mathbf{\Psi}_{s,\epsilon}$ could correspond with an s-specific choice of basis functions for a parameterization of $\mathbf{\Psi}$. The cross-validated ϵ-net estimation procedure will then select the basis (i.e., s) adaptively to the underlying truth ψ_0. In addition, various $\mathbf{\Psi}_s$ could represent guessed parametric or semi-parametric true subspaces of $\mathbf{\Psi}$.

Though the construction of ϵ-nets sounds abstract, a specific ϵ-net sieve can be obtained as follows, given a parameterization of $\mathbf{\Psi}$ in terms of a finite or countable parameter β ranging over a particular Euclidean set $\mathcal{B}$. Let $\mathbf{\Psi}^*$ denote one of the subspaces $\mathbf{\Psi}_s$. First, let the components β_j of the coefficient vector β be integer multiples of a given $\delta > 0$, then intersect the resulting discrete set with the original coefficient space $\mathcal{B}$. This yields a finite sieve of *δ-grids* $(\mathbf{\Psi}_\delta : \delta)$, indexed by $\delta > 0$, which corresponds with a sequence of *ϵ-nets*, indexed by $\epsilon > 0$. That is, there exists a function $\delta : \epsilon \to \delta(\epsilon)$, so that for all $\epsilon > 0$, $\mathbf{\Psi}_{\delta(\epsilon)}$ is an ϵ-net of $\mathbf{\Psi}^*$. The proposed candidate estimators for the resulting ϵ-net sieve simply minimize the empirical risk over the finite sets $\mathbf{\Psi}_\delta$. The resolution δ is then chosen with cross-validation. Consequently, the user does not need to know the actual function $\delta : \epsilon \to \delta(\epsilon)$ (i.e., how the ϵ-net resolution or approximation error ϵ depends on the choice of resolution δ for the δ-grid). The construction of an equally-spaced ϵ-net (i.e., an ϵ-net of approximately minimal size) can be much more challenging. However, if $D(\mathcal{S})$ is a Hilbert space and we parameterize $\mathbf{\Psi}^*$ with linear combinations of orthonormal basis functions, then the standard, equally-spaced δ-grid does provide an equally-spaced ϵ-net. Section 2.2.2 provides details on the construction of the standard δ-grid introduced above, as well as more general definitions of δ-grids.

2.1.2 Construction of a minimal empirical risk candidate estimator for each ϵ-net

Given a subspace-resolution pair (s, ϵ) and the corresponding discrete ϵ-net $\mathbf{\Psi}_{s,\epsilon}$, we define the following *minimum empirical risk estimator*

$$\hat{\Psi}_{s,\epsilon}(P_n) \equiv \text{argmin}_{\psi \in \mathbf{\Psi}_{s,\epsilon}} \int L(o, \psi) dP_n(o). \tag{8}$$

We note that for each fixed (s, ϵ), $\hat{\Psi}_{s,\epsilon}(P_n)$ can be viewed as an estimator of the *true risk minimizer* $\Psi_{s,\epsilon}(P_0)$ for the ϵ-net $\mathbf{\Psi}_{s,\epsilon}$, that is,

$$\Psi_{s,\epsilon}(P_0) \equiv \text{argmin}_{\psi \in \mathbf{\Psi}_{s,\epsilon}} \int L(o, \psi) dP_0(o). \tag{9}$$

2.1.3 Cross-validation selection of subspace $\mathbf{\Psi}_s$ and ϵ-net resolution ϵ

The subspace-resolution pair (s, ϵ) is selected with cross-validation as follows. Let $B_n \in \{0,1\}^n$ be a random binary n-vector whose observed value defines a split of the data $O_1, \ldots, O_n$, also called *learning sample*, into a validation sample and a training sample. If $B_n(i) = 0$, then observation i is placed in the *training sample*, and if $B_n(i) = 1$, it is placed in the *validation sample*. The choice of distribution for B_n corresponds with different cross-validation schemes, such as V-fold cross-validation and Monte-Carlo cross-validation. Denote the empirical distribution of the training and validation samples with P^0_{n,B_n} and P^1_{n,B_n}, respectively. The proportion of observations in the validation sample is denoted with $p \equiv \sum_i B_n(i)/n$. Bootstrap-based cross-validation can also be included in this framework by defining $B_n(i)$ as the number of times observation i occurs in the bootstrap sample. Then, the validation sample distribution, P^1_{n,B_n}, is the empirical distribution of all O_i with $B_n(i) = 0$, and the training sample distribution, P^0_{n,B_n}, is the weighted empirical distribution of the remaining observations, with weights being the counts $B_n(i)$.

For a given subspace $\mathbf{\Psi}_s$, let $\epsilon_n(k,s)$, $k = 1, \ldots, K_{2s}(n)$, be the set of values for the ϵ-net resolution ϵ. Let

$$\mathcal{A}_n \equiv \cup_{s=1}^{K_1(n)} \{(s, \epsilon_n(k,s)) : k = 1, \ldots, K_{2s}(n)\} \tag{10}$$

be the entire set of values considered for the subspace-resolution pairs (s, ϵ), and let

$$K_0(n) \equiv |\mathcal{A}_n| = \sum_{s=1}^{K_1(n)} K_{2s}(n), \tag{11}$$

be the cardinality of $\mathcal{A}_n$. The *cross-validation selector* $(s(P_n), \epsilon(P_n))$ of (s, ϵ) is defined as the minimizer of the cross-validated risk,

$$(s(P_n), \epsilon(P_n)) \equiv \text{argmin}_{(s,\epsilon)\in\mathcal{A}_n} E_{B_n} \int L(o, \hat{\Psi}_{s,\epsilon}(P^0_{n,B_n}))dP^1_{n,B_n}(o), \tag{12}$$

and the corresponding adaptive *cross-validated ϵ-net estimator* as

$$\hat{\Psi}(P_n) \equiv \hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n). \tag{13}$$

The contribution of the total number $K_0(n)$ of candidate estimators to the finite sample bound for the expected risk difference in Theorem 1 is $O(\log(K_0(n))/n)$.

Thus, if $K_0(n) = O(n^m)$, $m < \infty$, the contribution of $K_0(n)$ to the bound is $O(\log(n)/n)$ and is typically second order.

Implementation of the proposed cross-validated ϵ-net estimator $\hat{\Psi}(P_n)$ requires the construction of one or more ϵ-net sieves, $\{\mathbf{\Psi}_{s,\epsilon} : \epsilon\}$, and an algorithm for minimizing the empirical risk over all points in a given ϵ-net, $\mathbf{\Psi}_{s,\epsilon}$. These two issues are addressed next.

2.2 Construction of an ϵ-net sieve for a given parameter space

The following construction applies to each of the parameter spaces $\mathbf{\Psi}_s$, $s = 1, \ldots, K_1(n)$. However, to simplify notation, we suppress the index s and let $\mathbf{\Psi}^*$ denote one of the $\mathbf{\Psi}_s$ for a given choice s. The basis functions ϕ_j used to parameterize the parameter space $\mathbf{\Psi}^*$ are thus s-specific as well.

2.2.1 Parameterization of the parameter space

Given a parameter space $\mathbf{\Psi}^*$ and associated dissimilarity function d, let $\{\phi_j : j \in I\}$ be a countable collection of *basis functions*, so that each $\psi \in \mathbf{\Psi}^*$ can be arbitrarily well approximated by a known function g_0 (e.g., $g_0(x) = x$, $g_0(x) = 1/(1 + \exp(-x))$, or $g_0(x) = \exp(x)$) of finite linear combinations of such basis functions. For notational convenience, let $g_0(x) \equiv x$, and observe that, by redefining the parameter of interest as $g_0^{-1}(\psi_0)$, it suffices to work with linear approximations of $\mathbf{\Psi}^*$. Thus, for each $\psi \in \mathbf{\Psi}^*$, there exists a countable *index set* I and a corresponding vector of *coefficients* $(\beta_j : j \in I)$, so that $\psi = \sum_{j \in I} \beta_j \phi_j$. Hence, one can express the space $\mathbf{\Psi}^*$ as

$$\mathbf{\Psi}^* = \left\{ \sum_{j \in I} \beta_j \phi_j : \beta \in \mathcal{B} \subseteq \mathbb{R}^{|I|} \right\}, \tag{14}$$

where the *coefficient space* $\mathcal{B}$ is the Euclidean set defined as

$$\mathcal{B} \equiv \left\{ \beta \in \mathbb{R}^{|I|} : \sum_{j \in I} \beta_j \phi_j \in \mathbf{\Psi}^* \right\}. \tag{15}$$

It is assumed that $\mathcal{B}$ is a *bounded* Euclidean set, which one would expect to hold under Assumption A1 of Theorems 1 and 3: that is, to obtain a uniformly bounded loss function, one needs to bound the coefficient space $\mathcal{B}$.

Let $\psi_\beta \equiv \sum_{j \in I} \beta_j \phi_j$, so that we have the following parameterization of the parameter space: $\mathbf{\Psi}^* = \{\psi_\beta : \beta \in \mathcal{B}\}$.

Orthonormalization of basis functions. If the parameter space $\mathbf{\Psi}^*$ is embedded in a Hilbert space, one could orthonormalize the basis functions (in case the basis we start out with is not orthonormal) before constructing the ϵ-nets. The standard, equally-spaced δ-grid defined in Example 1, below, for the coefficient space $\mathcal{B}$, should then result in an equally-spaced (with respect to the Hilbert space) ϵ-net in the actual parameter space $\mathbf{\Psi}^*$.

Generalized basis functions. One could also specify a countable collection $\{\phi_{j,\alpha_j} : j\}$, of functions ϕ_{j,α_j}, indexed by a finite dimensional parameter α_j. For example, one could define $\phi_{j,\alpha_j}(W) \equiv \phi_j(\alpha_j W)$, where ϕ_j is a fixed basis function and α_j a smoothing degree for the basis function. Alternatively, $\phi_{j,\alpha}$ could correspond with the jth basis function from a collection of bases indexed by a parameter α. Minimizing the empirical risk over finite linear combinations $\sum_j \beta_j \phi_{j,\alpha_j}$ of such generalized basis functions involves simultaneously minimizing over the coefficients β_j and the parameters α_j. Barron (1993) provides universal approximation bounds for superpositions $\sum_j \beta_j \phi(\alpha_j W)$, where ϕ is a sigmoid function. An interesting question is whether a restriction to collections (indexed by the subspace s) of (known functions of) *linear* approximations, as in our methodology described above, has a formal disadvantage relative to using collections of *non-linear* approximations such as neural networks.

2.2.2 Construction of δ-grids

Definition 3 (δ-grid) *Given a function $g : \mathcal{B} \to \mathbb{R}$, from the Euclidean coefficient space $\mathcal{B} \subset \mathbb{R}^{|I|}$ into the real line, define a dissimilarity function d_g on $\mathcal{B}$ by*

$$d_g(\beta_1, \beta_2) \equiv |g(\beta_1) - g(\beta_2)|, \qquad \beta_1,\, \beta_2 \in \mathcal{B}. \tag{16}$$

For a given $\delta > 0$, a δ-grid with respect to the function g *is defined as a finite subset $\mathcal{B}_g(\delta)$ of $\mathcal{B}$ that satisfies the following two requirements.*
Condition Δ1. *For each $\beta \in \mathcal{B}_g(\delta)$ and coordinate $j \in \{1, \ldots, |I|\}$,*

$$\min_{\{\epsilon \neq 0:\ (\beta + \epsilon \vec{e}_j) \in \mathcal{B}_g(\delta)\}} d_g(\beta, \beta + \epsilon \vec{e}_j) = \delta, \tag{17}$$

where $\vec{e}_j$ denotes the unit $|I|$-vector with one in position j and zero elsewhere, $j \in \{1, \ldots, |I|\}$.

Condition Δ2. *For each $\beta \in \mathcal{B}_g(\delta)$ and coordinate $j \in \{1, \ldots, |I|\}$, if there exists $\epsilon \neq 0$ such that $d_g(\beta, \beta + \epsilon\vec{e}_j) = \delta$ and $(\beta + \epsilon\vec{e}_j) \in \mathcal{B}$, then $(\beta + \epsilon\vec{e}_j) \in \mathcal{B}_g(\delta)$.*

Condition Δ1 enforces the *grid structure* on the set $\mathcal{B}_g(\delta)$, in the sense that for any two neighboring points β_1 and β_2 in $\mathcal{B}_g(\delta)$, which only differ in one coordinate, the dissimilarity $d_g(\beta_1, \beta_2)$ equals δ. Condition Δ2 concerns the *richness* of the set $\mathcal{B}_g(\delta)$, that is, any neighboring point $\beta_2 \in \mathcal{B}$ of $\beta_1 \in \mathcal{B}_g(\delta)$, which only differs in one coordinate from β_1 and satisfies $d_g(\beta_1, \beta_2) = \delta$, also belongs to the δ-grid.

Note that the definition of a δ-grid suggests a general approach for constructing such a set. Starting with one point $\beta \in \mathcal{B}$, for each coordinate $j \in \{1, \ldots, |I|\}$, increase/decrease ϵ until $d_g(\beta, \beta + \epsilon\vec{e}_j) = \delta$. Then, add $(\beta + \epsilon\vec{e}_j)$ to the set $\mathcal{B}_g(\delta)$. Iterate by applying this process to each point in the current $\mathcal{B}_g(\delta)$. Since the dissimilarity function d_g is defined as a difference of a function evaluation, the procedure truly generates a set with the properties of a regular grid: e.g., any path between any two points β_1 and $\beta_2 \in \mathcal{B}_g(\delta)$ is of length a multiple of δ, that is, $d_g(\beta_1, \beta_2) \propto \delta$.

The dissimilarity function $d_g(\cdot, \cdot)$, for the coefficient space $\mathcal{B}$, implies a dissimilarity function $d_g^*(\cdot, \cdot)$, for the parameter space $\mathbf{\Psi}^*$: $d_g^*(\psi_{\beta_1}, \psi_{\beta_2}) \equiv d_g(\beta_1, \beta_2)$. Hence, the δ-grid $\mathcal{B}_g(\delta)$ for $\mathcal{B}$ implies a δ-grid $\mathbf{\Psi}_g(\delta)$ for $\mathbf{\Psi}^*$:

$$\mathbf{\Psi}_g(\delta) \equiv \{\psi_\beta : \beta \in \mathcal{B}_g(\delta)\}. \tag{18}$$

Note that the δ-grid $\mathbf{\Psi}_g(\delta)$ is finite, since by definition $\mathcal{B}_g(\delta)$ is finite. In addition, for each $\epsilon > 0$, there exists a $\delta(\epsilon)$ so that $\mathbf{\Psi}_g(\delta(\epsilon))$ is an ϵ-net of $\mathbf{\Psi}^*$. Thus, the δ-grid sieve $(\mathbf{\Psi}_g(\delta) : \delta)$ yields an ϵ-net sieve of $\mathbf{\Psi}^*$.

Below, we provide two examples of functions g specifying δ-grid pairs $\mathcal{B}_g(\delta)$ and $\mathbf{\Psi}_g(\delta)$, for the coefficient space $\mathcal{B}$ and parameter space $\mathbf{\Psi}^*$, respectively.

Example 1 (Standard, equally-spaced δ-grid) Consider the function $g : \mathcal{B} \to \mathbb{R}$ defined by:

$$g(\beta) \equiv \sum_{j \in I} \beta_j. \tag{19}$$

The δ-grid $\mathcal{B}_g(\delta)$, corresponding to this choice of function g applied to the coefficient space $\mathcal{B}$, has the form

$$\mathcal{B}_g(\delta) \equiv \{\beta \in \mathcal{B} : \beta_j/\delta \text{ is an integer for each coordinate } j \in I\}. \quad (20)$$

An algorithm is provided in Section 2.3, below, for minimizing the empirical risk over such a standard, equally-spaced δ-grid for $\mathcal{B}$.

Example 2 (Loss-based δ-grid) Consider the function $g : \mathcal{B} \to \mathbb{R}$, defined as the risk for the loss function L with respect to a known, possibly data-dependent, distribution $P \in \mathcal{M}$:

$$g(\beta) \equiv \int L(o, \psi_\beta) dP(o). \quad (21)$$

Our motivation for considering such loss-based functions g is the observation that standard, equally-spaced δ-grids for the coefficient space $\mathcal{B}$ (as in Example 1, above), based on a highly non-orthogonal parameterization $\{\psi_\beta : \beta \in \mathcal{B}\}$ of $\mathbf{\Psi}^*$, could result in poor (i.e., large size) ϵ-nets. The finite sample bound for the expected risk difference of the cross-validated ϵ-net estimator (Theorem 1, equation (26)) only depends on the ϵ-net through its approximation error, $B_0(s, \epsilon) = \min_{\psi \in \mathbf{\Psi}_{s,\epsilon}} \int (L(o, \psi) - L(o, \psi_0)) dP_0(o)$, and its cardinality, $N_s(\epsilon) = \mid \mathbf{\Psi}_{s,\epsilon} \mid$. This inspires us to define an ϵ-net sieve in terms of non-equally-spaced δ-grids for $\mathcal{B}$, so that two neighboring points in a given grid differ in empirical risk by δ. That is, we take the point of view in which the resolution for the space $\mathcal{B}$ is measured by the *empirical risk function*

$$g_n(\beta) \equiv \int L(o, \psi_\beta) dP_n(o) = \frac{1}{n} \sum_{i=1}^{n} L(O_i, \psi_\beta), \quad (22)$$

corresponding to the choice $P = P_n$, where P_n denotes the empirical distribution function. Thus,

$$\mathbf{\Psi}_{g_n}(\delta) \equiv \{\psi_\beta : \beta \in \mathcal{B}_{g_n}(\delta)\} \quad (23)$$

now represents a data-adaptive loss-based δ-grid for the parameter space $\mathbf{\Psi}^*$. The algorithm in Section 2.3, below, can be trivially applied to search over the non-equally-spaced grid $\mathcal{B}_{g_n}(\delta)$ for the coefficient space $\mathcal{B}$. In this context, dense regions of the grid for the coefficient space $\mathcal{B}$ correspond to a rapidly changing empirical risk surface, i.e., a surface with dense contours.

However, since the resulting ϵ-nets $\mathbf{\Psi}_{g_n}(\delta(\epsilon))$ now depend on the empirical distribution P_n, one cannot immediately apply Theorems 1 and 3. We plan to investigate in a simulation study the performance of the standard and data-adaptive loss-based δ-grid constructions of ϵ-nets.

Continuous sieve approximation of $\mathbf{\Psi}^*$. One can construct δ-grids as above for either the entire parameter space $\mathbf{\Psi}^*$ or a continuous sieve approximation of $\mathbf{\Psi}^*$ obtained as follows. Given $\epsilon > 0$, let $I_\epsilon \subset I$ be a finite index set of size $|I_\epsilon|$, so that the corresponding finite set $\{\phi_j : j \in I_\epsilon\}$ of basis functions generates an *ϵ-approximation* of $\mathbf{\Psi}^*$. That is, for each $\psi \in \mathbf{\Psi}^*$, we have $\inf_{\beta \in \mathcal{B}_\epsilon} d(\psi, \sum_{j \in I_\epsilon} \beta_j \phi_j) \leq \epsilon$, for the Euclidean set $\mathcal{B}_\epsilon \equiv \{(\beta_j : j \in I_\epsilon) : \beta \in \mathcal{B}\} \subseteq \mathbb{R}^{|I_\epsilon|}$. Let $\mathbf{\Psi}_\epsilon \equiv \{\psi_\beta : \beta \in \mathcal{B}_\epsilon\}$ be the corresponding element of the continuous sieve $(\mathbf{\Psi}_\epsilon : \epsilon)$ indexed by $\epsilon > 0$.

Constructing g-specific δ-grids for subspaces $\mathbf{\Psi}_\epsilon$, yields δ-grid pairs $\mathcal{B}_g(\delta, \epsilon)$ and $\mathbf{\Psi}_g(\delta, \epsilon)$, indexed by the parameter pair (δ, ϵ), for the grid resolution δ and resolution ϵ for the approximation $\mathbf{\Psi}_\epsilon$ of the space $\mathbf{\Psi}^*$. Note that each δ-grid $\mathbf{\Psi}_g(\delta, \epsilon)$ is finite, since by definition $\mathcal{B}_g(\delta, \epsilon)$ is finite. In addition, for each $\epsilon > 0$, there exists a $\delta(\epsilon)$ so that $\mathbf{\Psi}_g(\delta(\epsilon), \epsilon)$ is a 2ϵ-net of $\mathbf{\Psi}^*$. Thus, $(\mathbf{\Psi}_g(\delta, \epsilon) : (\delta, \epsilon))$ yields an ϵ-net sieve indexed by the pair (δ, ϵ), which is more flexible than sieves indexed by δ alone. In general, an ϵ-net sieve can be indexed by various additional parameters. The theorems in Section 3 show that as long as $\log(K(n))/n$ is of second order, where $K(n)$ is the total number of parameter values selected with cross-validation, an increase in the number of available finite subsets will generally improve the asymptotic performance of the resulting estimator (i.e., will increase its adaptivity to the true value ψ_0).

2.3 Algorithm for minimizing the empirical risk over an ϵ-net

Let $\mathcal{B}(\delta)$ and $\mathbf{\Psi}(\delta) = \{\psi_\beta = \sum_{j \in I} \beta_j \phi_j : \beta \in \mathcal{B}(\delta)\}$ denote a δ-grid pair, as defined in equation (20), where we omit the subscript g to simplify notation. Consider a function $f : \mathbf{\Psi}(\delta) \to \mathbb{R}$, which maps a parameter value $\psi \in \mathbf{\Psi}(\delta)$ into the empirical risk $\int L(o, \psi) dP_n(o)$. In order to compute the cross-validated ϵ-net estimator, we need an algorithm for minimizing such an empirical risk function f. Firstly, under the assumption that the coefficient space $\mathcal{B}$ is bounded, note that each $\beta \in \mathcal{B}(\delta)$ can be identified by an element in the lattice $\{0, \pm 1, \ldots, \pm M\}^{|I|}$, where $M = M(\delta)$ is a finite integer. Let

$f^* : \{0, \pm 1, \ldots, \pm M\}^{|I|} \to \mathbb{R}$ be an extension of the empirical risk function f to this lattice, where we set $f^*(x) \equiv \infty$ for any x which does not correspond with a point in $\mathcal{B}(\delta)$. We propose the following simple algorithm for minimizing f^* on the lattice $\{0, \pm 1, \ldots, \pm M\}^{|I|}$.

Initialize Set $k = 0$ and $x_k = (0, 0, \ldots, 0)$, an $|I|$-vector of zeros.

Define moves For any $x \in \{0, \pm 1, \ldots, \pm M\}^{|I|}$, let $\mathcal{S}(x)$ be the set of $2|I|$ vectors obtained by adding 1 to or subtracting 1 from a particular component x_j of the $|I|$-vector x. In case such a move results in a parameter value ψ outside the parameter space $\mathbf{\Psi}(\delta)$, one should consider a set of alternative moves (e.g., if adding 1 to x_j results in a parameter value outside the parameter space, then one can set any of the non-zero components of x equal to zero).

Iterate Let

$$x^* \equiv \text{argmin}_{x \in \mathcal{S}(x_k)} f^*(x).$$

If $f(x^*) \leq f(x_k)$, then set $k = k + 1$, $x_k = x^*$, and repeat. Otherwise, stop.

Output Let the final x^* be the candidate for the global minimum.

Starting values One could run this algorithm with various starting values. In particular, one could choose as starting values $\mathbf{\Psi}(\delta)$-discretized versions of initial available estimators, such as a penalized least squares estimator in the context of linear regression.

Obviously, users can define their own set of moves, depending on the particular parameter space they are minimizing over.

3 Finite sample results and implications

We prove two main theorems, which provide finite sample bounds for the expected risk difference between our proposed cross-validated ϵ-net estimator $\hat{\Psi}(P_n) = \hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n)$ and the parameter value $\Psi(P_0) = \psi_0$. Theorem 1 makes two fundamental assumptions: Assumption A1, that the loss function is uniformly bounded, and Assumption A2, that the variance of the ψ_0-centered loss function $L(O, \psi) - L(O, \psi_0)$ can be bounded by its expectation

uniformly in ψ. Theorem 3 only makes Assumption A1. By carrying out a Taylor series expansion, it can be informally argued that Assumption A2 can be expected to hold for loss functions satisfying the following property. Given any one-dimensional parametric submodel $\{P_{0,\epsilon} : \epsilon\}$, going through P_0 at $\epsilon = 0$ and with score in the Hilbert space $L_0^2(P_0)$ at $\epsilon = 0$,

$$\frac{d}{d\epsilon} \int (L(o, \Psi(P_{0,\epsilon})) - L(o, \Psi(P_0))) \, dP_0(o) \Big|_{\epsilon=0} = 0. \tag{24}$$

It is easy to verify that property (24) holds for the quadratic and negative log-density loss functions, which are two examples of loss functions satisfying Assumption A2. For this reason, we refer to a loss function satisfying Assumption A2 as a *quadratic loss function*. However, we stress that we have not established any formal equivalence between Assumption A2 and the identity in equation (24), and we do not conjecture such a formal equivalence. In a personal communication, Andrew Barron pointed out that the proofs below rely on similar empirical process techniques as those used in Barron (1991) for establishing finite sample inequalities for penalized minimum empirical risk estimators.

3.1 Finite sample inequality for quadratic loss functions

Let

$$\begin{aligned} B_0(s, \epsilon) &= B(s, \epsilon \mid P_0) \equiv \min_{\psi \in \mathbf{\Psi}_{s,\epsilon}} d_0(\psi, \psi_0) \\ &= \min_{\psi \in \mathbf{\Psi}_{s,\epsilon}} \int (L(o, \psi) - L(o, \psi_0)) \, dP_0(o) \\ &= \int (L(o, \Psi_{s,\epsilon}(P_0)) - L(o, \Psi(P_0))) \, dP_0(o) \end{aligned} \tag{25}$$

denote the *risk approximation error* or *risk resolution* of the ϵ-net $\mathbf{\Psi}_{s,\epsilon}$, where $\Psi_{s,\epsilon}(P_0)$, defined in equation (9), is the true risk minimizer for this ϵ-net. In the following theorem, we establish a bound on the expectation of the random (via the empirical distribution P_n) risk difference:

$$E_{B_n} d_0(\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n}), \psi_0) = E_{B_n} \int \Big(L(o, \hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n})) - L(o, \psi_0) \Big) \, dP_0(o).$$

Theorem 1 Quadratic loss functions.
Assumptions.
A1. *There exists an* $M_1 < \infty$ *so that*

$$\sup_{\psi \in \mathbf{\Psi}} \sup_{O} \mid L(O, \psi) - L(O, \psi_0) \mid \leq M_1,$$

where the supremum is taken over a support of the distribution P_0 *of* O.
A2. *There exists an* $M_2 < \infty$ *so that*

$$\sup_{\psi \in \mathbf{\Psi}} \frac{VAR_{P_0}[L(O, \psi) - L(O, \psi_0)]}{E_{P_0}[L(O, \psi) - L(O, \psi_0)]} \leq M_2.$$

Definitions. *Define the following constant*

$$C(\lambda) \equiv 2(1 + \lambda)^2 \left(\frac{2M_1}{3} + \frac{M_2}{\lambda} \right).$$

Finite sample result. *For any* $\lambda > 0$*, we have the following finite sample inequality for the expected risk difference between the proposed cross-validated adaptive* ϵ*-net estimator* $\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n)$ *and the parameter value* $\Psi(P_0) = \psi_0$*:*

$$\begin{aligned} & Ed_0(\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n}), \psi_0) \leq \\ & \quad (1 + 2\lambda) \min_{(s,\epsilon) \in \mathcal{A}_n} \left\{ (1 + 2\lambda) B_0(s, \epsilon) + 2C(\lambda) \frac{1 + \log(N_s(\epsilon))}{n(1 - p)} \right\} \\ & \quad + 2C(\lambda) \frac{1 + \log(K_0(n))}{np}. \end{aligned} \tag{26}$$

Recall that $K_0(n)$ is the total number of candidate minimum empirical risk estimators, $\hat{\Psi}_{s,\epsilon}$, indexed by the subspace-resolution pairs (s, ϵ), and that $N_s(\epsilon)$ is the cardinality of the ϵ-net $\mathbf{\Psi}_{s,\epsilon}$, for such a choice of subspace $\mathbf{\Psi}_s$ and resolution ϵ.

3.2 Proof of Theorem 1

The proof of Theorem 1 involves a double application of the following general theorem, which compares the risk of the estimator chosen by the cross-validation selector $k(P_n)$, with the risk of the estimator chosen with an oracle selector $\tilde{k}(P_n)$, among a set of candidate estimators $\{\hat{\psi}_k = \hat{\Psi}_k(P_n) : k = 1, \ldots, K(n)\}$. In the first application of Theorem 2, $K(n)$ refers to the total number $K_0(n)$ of subspace-resolution pairs (s, ϵ); in the second application, $K(n)$ refers to the size $N_s(\epsilon)$ of an ϵ-net $\mathbf{\Psi}_{s,\epsilon}$ for a given pair (s, ϵ).

Theorem 2 *Let* $\{\hat{\psi}_k = \hat{\Psi}_k(P_n) : k = 1, \ldots, K(n)\}$ *be a given set of* $K(n)$ *candidate estimators of the parameter value* $\psi_0 = argmin_{\psi \in \mathbf{\Psi}} \int L(o, \psi) dP_0(o)$. *Suppose that* $\hat{\Psi}_k(P_n) \in \mathbf{\Psi}$ *for all* k, *with probability 1. Let* $k(P_n) \equiv argmin_k E_{B_n} \int L(o, \hat{\Psi}_k(P^0_{n,B_n})) dP_0(o)$ *be the* cross-validation selector, *and let* $\tilde{k}(P_n) \equiv argmin_k E_{B_n} \int L(o, \hat{\Psi}_k(P^0_{n,B_n})) dP_0(o)$ *be the comparable* benchmark *or* oracle selector. *Then, under Assumptions A1 and A2 of Theorem 1, one has the following finite sample inequality, for any* $\lambda > 0$*:*

$$Ed_0(\hat{\Psi}_{k(P_n)}(P^0_{n,B_n}), \psi_0) \leq (1 + 2\lambda) Ed_0(\hat{\Psi}_{\tilde{k}(P_n)}(P^0_{n,B_n}), \psi_0) + 2C(\lambda) \frac{1 + \log(K(n))}{np}. \quad (27)$$

The reader is referred to Theorem 2 in Dudoit and van der Laan (2003) for a proof of this result. An alternative, shorter proof is provided in van der Vaart (2003). van der Vaart (2003) also provides extensions of this result to unbounded loss functions, with conditions on the tail of the distribution of the loss function. Theorem 2 above is a special case of the general Theorem 1 for quadratic loss functions in van der Laan and Dudoit (2003).

Proof of Theorem 1. The proof is a double application of Theorem 2. The first, or *outer*, application of Theorem 2 concerns cross-validation selection among the $K_0(n)$ minimum empirical risk estimators $\hat{\Psi}_{s,\epsilon}(P_n)$, indexed by subspace-resolution pairs $(s, \epsilon) \in \mathcal{A}_n$. The second, or *inner*, application of the theorem concerns selection among the $N_s(\epsilon)$ candidate values in a particular ϵ-net, $\mathbf{\Psi}_{s,\epsilon}$. In the latter case, we use the fact that for constant (i.e., non-random) estimators, the cross-validated risk equals the empirical risk. Thus, the outer application yields the $\frac{1+\log(K_0(n))}{np}$ term, while the inner application yields the $\frac{1+\log(N_s(\epsilon))}{n(1-p)}$ term, in the finite sample inequality of equation (26).

Outer application of Theorem 2. First apply Theorem 2 to the candidate estimators $\hat{\Psi}_k(P_n) \equiv \hat{\Psi}_{s(k),\epsilon(k)}(P_n)$, where k indexes the subspace-resolution pairs (s, ϵ) in the set of possible values $\mathcal{A}_n = \{(s(k), \epsilon(k)) : k = 1, \ldots, K_0(n)\}$. Note that $\hat{\Psi}_{k(P_n)}(P_n) = \hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n)$ denotes the cross-validated ϵ-net es-

timator. Theorem 2 yields the following inequality:

$$\begin{aligned} Ed_0(\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n}),\psi_0) &\leq (1+2\lambda)Ed_0(\hat{\Psi}_{\tilde{k}(P_n)}(P^0_{n,B_n}),\psi_0) \\ &\quad + 2C(\lambda)\frac{1+\log(K_0(n))}{np}. \end{aligned}$$

The main term on the right-hand side can be rewritten and bounded as follows:

$$\begin{aligned} Ed_0(\hat{\Psi}_{\tilde{k}(P_n)}(P^0_{n,B_n}),\psi_0) &= E\int \Big(L(o,\hat{\Psi}_{\tilde{k}(P_n)}(P^0_{n,B_n})) - L(o,\psi_0)\Big)\,dP_0(o) \\ &= E\min_{k\in\{1,\ldots,K_0(n)\}}\int \Big(L(o,\hat{\Psi}_k(P^0_{n,B_n})) - L(o,\psi_0)\Big)\,dP_0(o) \\ &\leq \min_{k\in\{1,\ldots,K_0(n)\}} E\int \Big(L(o,\hat{\Psi}_k(P^0_{n,B_n})) - L(o,\psi_0)\Big)\,dP_0(o) \\ &= \min_{(s,\epsilon)\in\mathcal{A}_n} E\int \Big(L(o,\hat{\Psi}_{s,\epsilon}(P^0_{n,B_n})) - L(o,\psi_0)\Big)\,dP_0(o) \quad (28) \end{aligned}$$

Inner application of Theorem 2. For each fixed subspace-resolution pair (s,ϵ), we now apply Theorem 2 with non-random candidate "estimators", $\psi_k^{s,\epsilon}$, $k=1,\ldots,N_s(\epsilon)$, corresponding to the points in the ϵ-net $\mathbf{\Psi}_{s,\epsilon}$. In this application, the empirical distribution is P^0_{n,B_n} and corresponds to a particular training sample of size $n(1-p)$ for the cross-validation selection of (s,ϵ). For notational convenience, however, we apply Theorem 2 with an empirical distribution P_n and substitute P^0_{n,B_n} for P_n in the resulting finite sample inequality. Thus, let $\hat{\Psi}_k(P_n) \equiv \psi_k^{s,\epsilon}$ and let B_n^* denote the binary split vector defining the cross-validation scheme. Because the candidate estimators $\hat{\Psi}_k(P_n)$ are constant, we have, for any B_n^*,

$$\begin{aligned} k(P_n) &= \text{argmin}_{k\in\{1,\ldots,N_s(\epsilon)\}} E_{B_n^*}\int L(o,\hat{\Psi}_k(P^0_{n,B_n^*}))dP^1_{n,B_n^*}(o) \\ &= \text{argmin}_{k\in\{1,\ldots,N_s(\epsilon)\}}\int L(o,\psi_k^{s,\epsilon})dP_n(o), \end{aligned}$$

which shows that $\hat{\Psi}_{k(P_n)}(P_n) = \hat{\Psi}_{s,\epsilon}(P_n)$, where $\hat{\Psi}_{s,\epsilon}(P_n)$ is the (s,ϵ)-specific minimum empirical risk estimator for the ϵ-net $\mathbf{\Psi}_{s,\epsilon}$ (equation (8)). To summarize, in this setting of non-random candidate estimators, the quantities in

Theorem 2 have the following analogues

$$\begin{aligned}
\hat{\Psi}_k(P_n) &= \psi_k^{s,\epsilon}, \\
\hat{\Psi}_{k(P_n)}(P^0_{n,B_n^*}) &= \hat{\Psi}_{k(P_n)}(P_n) = \hat{\Psi}_{s,\epsilon}(P_n), \\
\hat{\Psi}_{\tilde{k}(P_n)}(P_n) &= \Psi_{s,\epsilon}(P_0),
\end{aligned}$$

where $\Psi_{s,\epsilon}(P_0)$ is true risk minimizer for the ϵ-net defined in equation (9). Thus, application of Theorem 2 with these analogues gives us the following finite sample inequality, for any $\lambda > 0$,

$$\begin{aligned}
E \int L^*(o, \hat{\Psi}_{s,\epsilon}(P_n)) dP_0(o) &\leq (1 + 2\lambda) \int L^*(o, \Psi_{s,\epsilon}(P_0)) dP_0(o) \\
&\quad + 2C(\lambda) \frac{1 + \log(N_s(\epsilon))}{np^*},
\end{aligned}$$

where we use the short-hand notation $L^*(O, \psi) \equiv L(O, \psi) - L(O, \psi_0)$. Since this inequality can be applied for any cross-validation scheme (i.e., any distribution for the binary split vector B_n^*), we can set $p^* = 1$ to achieve the sharpest bound. Thus, for any $\lambda > 0$,

$$\begin{aligned}
E \int L^*(o, \hat{\Psi}_{s,\epsilon}(P_n)) dP_0(o) &\leq (1 + 2\lambda) \int L^*(o, \Psi_{s,\epsilon}(P_0)) dP_0(o) \\
&\quad + 2C(\lambda) \frac{1 + \log(N_s(\epsilon))}{n}.
\end{aligned}$$

Finally, application of this inequality to the empirical distribution P^0_{n,B_n}, corresponding to a training sample of size $n(1-p)$ for a given B_n, yields:

$$\begin{aligned}
E_{|B_n} \int L^*(o, \hat{\Psi}_{s,\epsilon}(P^0_{n,B_n})) dP_0(o) &\leq (1 + 2\lambda) \int L^*(o, \Psi_{s,\epsilon}(P_0)) dP_0(o) \\
&\quad + 2C(\lambda) \frac{1 + \log(N_s(\epsilon))}{n(1-p)}.
\end{aligned}$$

Since the bound on the right-hand side does not depend on the split vector B_n, it also holds unconditionally. Substituting this last bound into equation (28) and noting that, by definition, $B_0(s, \epsilon) = \int L^*(o, \Psi_{s,\epsilon}(P_0)) dP_0(o)$, yields the reported finite sample bound in equation (26). This completes the proof of Theorem 1.

□

3.3 Adaptivity

The *covering number* $N(\epsilon, \mathbf{\Psi}_s, d)$ of the parameter space $\mathbf{\Psi}_s$ is defined as the minimal number of points needed to obtain an ϵ-net $\mathbf{\Psi}_{s,\epsilon}$ of $\mathbf{\Psi}_s$ (examples of covering numbers are provided in Section 3.4, below). Consequently, the ϵ-net size $N_s(\epsilon)$ can always be chosen to be of the same order as $N(\epsilon, \mathbf{\Psi}_s, d)$. Thus, the finite sample inequality in Theorem 1 proves that the risk of the cross-validated adaptive ϵ-net estimator of ψ_0 converges to the optimal risk of ψ_0 at a rate as fast or faster than

$$r_{opt}(n) \equiv \max \left\{ \min_{(s,\epsilon) \in \mathcal{A}_n} \left\{ B_0(s, \epsilon) + \frac{\log(N(\epsilon, \mathbf{\Psi}_s, d))}{n} \right\}, \frac{\log(K_0(n))}{n} \right\}. \quad (29)$$

For loss functions satisfying Assumption A2 of Theorem 1 and dissimilarity functions d corresponding with a distance defined by a norm, one will typically have

$$\sup_{\{\psi : d(\psi, \psi_0) \leq \epsilon\}} \int (L(o, \psi) - L(o, \psi_0)) dP_0(o) \leq C\epsilon^2, \quad (30)$$

for some $C < \infty$. That is, the loss function is quadratic as defined by equation (24), above. The bound in equation (30) holds, for example, under Assumption A1 for the squared error and negative log-density loss functions, and with the L^2-distance or supremum norm distance d. Then, we have that the risk approximation error $B_0(s, \epsilon) \leq C\epsilon^2$, for any s for which the minimal distance between the ϵ-net $\mathbf{\Psi}_{s,\epsilon}$ and ψ_0 is less than or equal to ϵ. Thus, in this case, the rate of convergence $r_{opt}(n)$ can be bounded as follows:

$$r_{opt}(n) \leq \max \left\{ \min_{\{s : \psi_0 \in \mathbf{\Psi}_s\}} \min_{\epsilon} \left\{ \epsilon^2 + \frac{\log(N(\epsilon, \mathbf{\Psi}_s, d))}{n} \right\}, \frac{\log(K_0(n))}{n} \right\}. \quad (31)$$

Under the condition that $\log(K_0(n)) = O(\log(n))$, for infinite dimensional parameter spaces $\mathbf{\Psi}_s$, the first term within the max-operator will typically dominate. The above bound is a bound on the worst-case risk difference, since it does not depend anymore on the actual data generating distribution P_0. In Section 3.4, below, we verify that for well-known smoothness classes $\mathbf{\Psi}_s = \{\Psi(P) : P \in \mathcal{M}_s\}$, for multivariate real-valued functions, the above explicit bound

$$\min_{\epsilon} \left\{ \epsilon^2 + \frac{\log(N(\epsilon, \mathbf{\Psi}_s, \| \cdot \|_\infty))}{n} \right\} \quad (32)$$

corresponds with the *optimal minimax rate of convergence* defined as

$$\min_{\hat{\Psi}(P_n)} \max_{P_0 \in \mathcal{M}_s} d_0(\hat{\Psi}(P_n), \psi_0).$$

We refer to van der Vaart and Wellner (1996) for the covering numbers, $N(\epsilon, \mathbf{\Psi}_s, \|\cdot\|_\infty)$, of these classes of functions $\mathbf{\Psi}_s$ with respect to (w.r.t.) the supremum norm $\|\cdot\|_\infty$. Yang and Barron (1999) provide a theory showing that, in general, the main term in equation (29) corresponds with the optimal minimax rate of convergence for the parameter space $\mathbf{\Psi}_s$.

Since $r_{opt}(n)$ involves the minimum of these optimal minimax rates of convergence, over all subspaces $\mathbf{\Psi}_s$ containing the true ψ_0, this shows that the cross-validated adaptive ϵ-net estimator is indeed *adaptive*. That is, it achieves at worst the minimax rate of convergence corresponding with the smallest subspace $\mathbf{\Psi}_s$ containing the true parameter value ψ_0.

Since the risk approximation error $B_0(s, \epsilon)$ depends on the true parameter value ψ_0, $r_{opt}(n)$ could be significantly smaller than the above universal (distribution free) bound, which substitutes for $B_0(s, \epsilon)$ the upper bound ϵ^2 (for each s with $\psi_0 \in \mathbf{\Psi}_s$). To conclude, the cross-validated adaptive ϵ-net estimator $\hat{\Psi}(P_n)$ is indeed capable of adapting to actual properties of ψ_0 and, thereby, possibly achieves a better rate of convergence than the worst-case optimal rate implied by the size of the parameter space $\mathbf{\Psi}$.

3.4 Examples of covering numbers

Results on covering numbers $N(\epsilon, \mathbf{\Psi}_s, \|\cdot\|)$, w.r.t. the supremum norm or other norms, can be found in approximation theory. For example, Canuto and Quarteroni (1982) provide approximation results for polynomial sieves, which yield upper bounds for the ϵ-net size $N_s(\epsilon)$ for a multivariate Sobolev smoothness class $\mathbf{\Psi}_s$. We refer to van der Vaart and Wellner (1996), Section 2.7, for results and corresponding proofs regarding covering numbers w.r.t. the supremum norm, for various general classes of functions $\mathbf{\Psi}_s$. One of the general examples is summarized below.

Example 3 (Lipschitz functions on Euclidean sets, Theorem 2.7.1, van der Vaart and Wellner (1996)) Define, for any m-vector $k = (k_1, \ldots, k_m)$ of non-negative integers, the differential operator

$$D^k \equiv \frac{d^{k.}}{dx_1^{k_1} \ldots dx_m^{k_m}},$$

where $k_\cdot = \sum_i k_i$. For a positive real number α, let $\underline{\alpha} = \lfloor \alpha \rfloor$ be the largest integer less than or equal to α (i.e., floor). For a function $f : \mathcal{X} \subset \mathbb{R}^m \to \mathbb{R}$, let

$$\| f \|_\alpha \equiv \max_{k_\cdot \leq \alpha} \sup_x | D^k f(x) | + \max_{k_\cdot = \underline{\alpha}} \sup_{x,y} \frac{| D^k f(x) - D^k f(y) |}{\| x - y \|^{\alpha - \underline{\alpha}}}.$$

Here, the suprema are taken over all x, y in the interior of $\mathcal{X}$, with $x \neq y$. Let

$$C_M^\alpha \equiv \{ f : f : \mathcal{X} \to \mathbb{R}, \ \| f \|_\alpha \leq M \}, \tag{33}$$

where we assume that $\mathcal{X}$ is a bounded, convex subset of $\mathbb{R}^m$ with non-empty interior. There exists a constant K, depending only on α and m, such that for every $\epsilon > 0$,

$$\log(N(\epsilon, C_1^\alpha(\mathcal{X}), \| \cdot \|_\infty)) \leq K \lambda(\mathcal{X}_1) \left(\frac{1}{\epsilon} \right)^{m/\alpha},$$

where $\lambda(\mathcal{X}_1)$ is the Lebesgue measure of the set $\mathcal{X}_1 \equiv \{ x : \inf_{y \in \mathcal{X}} \| x - y \| < 1 \}$ and $\| \cdot \|_\infty$ denotes the supremum norm over $\mathcal{X}$. If $\mathbf{\Psi}_s = C_M^\alpha$, then

$$\min_\epsilon \left\{ \epsilon^2 + \frac{\log(N(\epsilon, \mathbf{\Psi}_s, \| \cdot \|_\infty))}{n} \right\} = O(n^{-\frac{2\alpha}{2\alpha + m}}),$$

which is the well-known minimax rate of convergence (pointwise and w.r.t. to L^2-norms) in regression and density estimation.

3.5 Finite sample inequality for general loss functions

For general loss functions $L(O, \psi)$, which are not required to satisfy Assumption A2 of Theorem 1, we have the following finite sample result for the expected risk difference between our proposed cross-validated ϵ-net estimator $\hat{\Psi}(P_n) = \hat{\Psi}_{s(P_n), \epsilon(P_n)}(P_n)$ and the parameter value $\Psi(P_0) = \psi_0$.

Theorem 3 General loss functions.
Assumption.
A1. *There exists an $M_1 < \infty$ so that*

$$\sup_{\psi \in \mathbf{\Psi}} \sup_O | L(O, \psi) - L(O, \psi_0) | \leq M_1,$$

where the supremum is taken over a support of the distribution P_0 of O.
Finite sample result. *We have the following finite sample inequality for*

the expected risk difference between the proposed cross-validated adaptive ϵ-net estimator $\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n)$ and the parameter value $\Psi(P_0) = \psi_0$:

$$Ed_0(\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n}), \psi_0) \leq \min_{(s,\epsilon)\in\mathcal{A}_n} \left\{ B_0(s,\epsilon) + 4M_1 \frac{\sqrt{\log N_s(\epsilon)}}{\sqrt{n(1-p)}} \right\} + 4M_1 \frac{\sqrt{\log K_0(n)}}{\sqrt{np}}. \quad (34)$$

Implications of this theorem, in terms of optimality and adaptivity for the cross-validated ϵ-net estimator, can be discussed in the same manner as above for quadratic loss functions and are therefore not repeated here.

3.6 Proof of Theorem 3

The proof of Theorem 3 is based on the following general theorem for the cross-validation selector, which is the general loss function analogue of Theorem 2 for quadratic loss functions.

Theorem 4 *Let $\{\hat{\psi}_k = \hat{\Psi}_k(P_n) : k = 1, \ldots, K(n)\}$ be a given set of $K(n)$ candidate estimators of the parameter value $\psi_0 = argmin_{\psi\in\Psi} \int L(o,\psi)dP_0(o)$. Suppose that $\hat{\Psi}_k(P_n) \in \Psi$ for all k, with probability 1. Let $k(P_n) \equiv argmin_k E_{B_n} \int L(o, \hat{\Psi}_k(P^0_{n,B_n}))dP_1$ be the* cross-validation selector*, and let $\tilde{k}(P_n) \equiv argmin_k E_{B_n} \int L(o, \hat{\Psi}_k(P^0_{n,B_n}))dP_0(o)$ be the comparable* benchmark *or* oracle selector*. Then, under Assumption A1 of Theorem 3, one has the following finite sample inequality:*

$$Ed_0(\hat{\Psi}_{k(P_n)}(P^0_{n,B_n}), \psi_0) \leq Ed_0(\hat{\Psi}_{\tilde{k}(P_n)}(P^0_{n,B_n}), \psi_0) + \frac{4M_1\sqrt{\log(K(n))}}{\sqrt{np}}. \quad (35)$$

Given Theorem 4, the proof of Theorem 3 is completely analogous to the proof of Theorem 1 for quadratic loss functions; it is therefore not repeated here.

Proof of Theorem 4. The following proof is from van der Vaart (2003). For notational convenience, let $L^*(O,\psi) \equiv L(O,\psi) - L(O,\psi_0)$. Firstly, observe that, by definition of the cross-validation selector $k(P_n)$,

$$E_{B_n} \int L^*(o, \hat{\Psi}_{k(P_n)}(P^0_{n,B_n}))dP^1_{n,B_n}(o) \leq E_{B_n} \int L^*(o, \hat{\Psi}_{\tilde{k}(P_n)}(P^0_{n,B_n}))dP^1_{n,B_n}(o).$$

This inequality can be rewritten in the form

$$E_{B_n} \int L^*(o, \hat{\Psi}_{k(P_n)}(P^0_{n,B_n}))dP_0(o) \leq E_{B_n} \int L^*(o, \hat{\Psi}_{\tilde{k}(P_n)}(P^0_{n,B_n}))dP_0(o)$$
$$+\frac{1}{\sqrt{np}} E_{B_n} \int \left(L^*(o, \hat{\Psi}_{\tilde{k}(P_n)}(P^0_{n,B_n})) - L^*(o, \hat{\Psi}_{k(P_n)}(P^0_{n,B_n})) \right) dG^1_{n,B_n}(o),$$

where $G^1_{n,B_n} \equiv \sqrt{np}(P^1_{n,B_n} - P_0)$ is the empirical process based on the np observations of the validation set (i.e., O_i with $B_n(i) = 1$). Next, one can split the integral on the right-hand side and replace the risk difference for the two randomly selected estimators (corresponding to $k(P_n)$ and $\tilde{k}(P_n)$) by a maximum over all $K(n)$ candidate estimators $\{\hat{\Psi}_k(P_n) : k = 1, \ldots, K(n)\}$. This gives us

$$E_{B_n} \int L^*(o, \hat{\Psi}_{k(P_n)}(P^0_{n,B_n}))dP_0(o) \leq E_{B_n} \int L^*(o, \hat{\Psi}_{\tilde{k}(P_n)}(P^0_{n,B_n}))dP_0(o)$$
$$+\frac{2}{\sqrt{np}} E_{B_n} \max_k \int L^*(o, \hat{\Psi}_k(P^0_{n,B_n}))dG^1_{n,B_n}(o).$$

Given a class of functions $\mathcal{F}$, let $N(\epsilon, \mathcal{F}, L^2(Q))$ be the minimal number of balls of size ϵ needed to cover $\mathcal{F}$ in the Hilbert space $L^2(Q)$. Formula (2.5.5) in van der Vaart and Wellner (1996) shows that

$$E \sup_{f \in \mathcal{F}} \mid G_n f \mid \leq \int_0^\infty \sqrt{\log \left(\sup_Q N(\epsilon \parallel F \parallel_{Q,2}, \mathcal{F}, L^2(Q)) \right)} d\epsilon \parallel F \parallel_{P_0,2},$$

where $G_n f \equiv \int f(o)\sqrt{n}(dP_n - dP_0)(o)$, $F \equiv \sup_{f \in \mathcal{F}} \mid f \mid$ is the envelope of $\mathcal{F}$, and $\parallel F \parallel_{P,2} \equiv \sqrt{\int F^2 dP}$. In particular, given a finite class $\mathcal{F}$ of functions of O,

$$E \max_{f \in \mathcal{F}} \mid G_n f \mid \leq C(\mathcal{F}) \sqrt{\log(|\mathcal{F}|)} \max_{f \in \mathcal{F}} \parallel f \parallel_\infty,$$

where $C(\mathcal{F}) \equiv \min\{\epsilon : \sup_Q N(\epsilon \parallel F \parallel_{Q,2}, \mathcal{F}, L^2(Q)) = 1\}$. If $\epsilon_1 > 2 \parallel F \parallel_{Q,2}$, then $N(\epsilon, \mathcal{F}, L^2(Q)) = 1$. This follows from the fact that for any f_1 and f_2 in $\mathcal{F}$, $\parallel f_1 - f_2 \parallel_{Q,2} \leq \parallel f_1 \parallel_{Q,2} + \parallel f_2 \parallel_{Q,2} \leq 2 \parallel F \parallel_{Q,2}$. Thus, $C(\mathcal{F}) = 2$. Applying this result to $\mathcal{F} = \{O \to L^*(O, \hat{\Psi}_k(P^0_{n,B_n})) : k = 1, \ldots, K(n)\}$ gives us

$$E \max_k \int L^*(o, \hat{\Psi}_k(P^0_{n,B_n}))dG^1_{n,B_n}(o) \leq 2\sqrt{\log K(n)} M_1.$$

This proves the theorem.

□

4 Applications to regression and density estimation

Given a sieve of finite sets, $\mathbf{\Psi}_{s,\epsilon} \subseteq \mathbf{\Psi}$, indexed by pairs $(s, \epsilon) \in \mathcal{A}_n$, let $\hat{\Psi}(P_n) = \hat{\Psi}_{s(P_n),\epsilon(P_n)}(P_n)$ be the cross-validated adaptive ϵ-net estimator as defined in equation (13). Application of general Theorem 1, yields the following finite sample inequalities for this estimator in multivariate regression and density estimation.

4.1 Univariate outcome regression

Corollary 1 Univariate outcome regression.
Setting. *Let $O = (W, Y) \sim P_0$, where Y is a scalar outcome and W is a vector of covariates with c.d.f. F_0. Consider the conditional mean outcome parameter $\psi_0(W) \equiv E_{P_0}[Y \mid W]$, with corresponding loss function the* quadratic loss function $L(O, \psi) \equiv (Y - \psi(W))^2$.
Assumptions. *Assume that there exists a constant $C_0 < \infty$, so that $|Y| \leq C_0$ a.s. and $\sup_{\psi \in \mathbf{\Psi}} \sup_W \mid \psi(W) \mid \leq C_0$.*
Definitions. *Let $M_1 \equiv 4C_0^2$ and $M_2 \equiv 16C_0^2$.*
Finite sample result. *The finite sample inequality of Theorem 1 holds with constants M_1 and M_2 as defined above. That is, for any $\lambda > 0$,*

$$
\begin{aligned}
E \int (\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n})(w) - \psi_0(w))^2 dF_0(w) \qquad (36)\\
\leq (1 + 2\lambda) \times \min_{(s,\epsilon) \in \mathcal{A}_n} \Bigg\{ (1 + 2\lambda) \min_{\psi \in \mathbf{\Psi}_{s,\epsilon}} \int (\psi(w) - \psi_0(w))^2 dF_0(w) \\
+ \; 2C(\lambda) \frac{1 + \log(N_s(\epsilon))}{n(1-p)} \Bigg\} \\
+ 2C(\lambda) \frac{1 + \log(K_0(n))}{np}.
\end{aligned}
$$

Proof of Corollary 1. First recall that the risk difference $d_0(\psi, \psi_0)$ equals the expected value of the squared difference between a candidate ψ and the truth ψ_0, that is,

$$
d_0(\psi, \psi_0) = \int (L(o, \psi) - L(o, \psi_0)) dP_0(o) = \int (\psi(w) - \psi_0(w))^2 dF_0(w).
$$

Application of Theorem 1 requires verification of Assumptions A1 and A2. Regarding Assumption A1, we note that

$$L(O,\psi) - L(O,\psi_0) = (Y - \psi(W))^2 - (Y - \psi_0(W))^2.$$

Thus, Assumption A1 holds with $M_1 = 4C_0^2$. Regarding Assumption A2, we note that

$$\begin{aligned} E_{P_0}\left[(L(O,\psi) - L(O,\psi_0))^2\right] &= \int \left((y-\psi(w))^2 - (y-\psi_0(w))^2\right)^2 dP_0(o) \\ &= \int (\psi(w) - \psi_0(w))^2 (2y - \psi(w) - \psi_0(w))^2 dP_0(o) \\ &\leq 16C_0^2 \int (\psi(w) - \psi_0(w))^2 dF_0(w) \\ &= 16C_0^2 \; E_{P_0}\left[L(O,\psi) - L(O,\psi_0)\right]. \end{aligned}$$

Thus, Assumption A2 holds with $M_2 = 16C_0^2$. This proves the corollary.

□

Example 4 (Linear regression: Adaptation to sparsity) Consider the regression setting as described in Corollary 1, where we assume that there exists a constant $C_0 < \infty$, so that $Pr(|Y| < C_0) = 1$. Given a set of basis functions ϕ_j, $j = 1, \ldots, N$, let $\mathbf{\Psi} \equiv \{\psi_\beta : \psi_\beta \equiv \sum_{j=1}^N \beta_j \phi_j, \; \sup_W \mid \sum_{j=1}^N \beta_j \phi_j(W) \mid < C_0\}$ be the parameter space. This corresponds with assuming that the regression function $E_{P_0}[Y \mid W]$ is linear in the basis functions ϕ_j. Let $\mathbf{\Psi}_k \equiv \{\psi_\beta \in \mathbf{\Psi} : \sum_{j=1}^N \mathrm{I}(\beta_j \neq 0) \leq k\} \subseteq \mathbf{\Psi}$ be the set of linear regression functions with maximally k non-zero coefficients, $k = 1, \ldots, N$. For each (k,ϵ)-pair, let $\mathbf{\Psi}_{k,\epsilon}$ be an ϵ-net of $\mathbf{\Psi}_k$, consisting of $N_k(\epsilon)$ points. Let $\hat{\Psi}_{k,\epsilon}(P_n)$ be the *empirical* risk minimizer over the ϵ-net $\mathbf{\Psi}_{k,\epsilon}$ and let $(k(P_n), \epsilon(P_n))$ denote the minimizer of the *cross-validated* risk

$$(k,\epsilon) \rightarrow E_{B_n} \int L(o, \hat{\Psi}_{k,\epsilon}(P^0_{n,B_n}))dP^1_{n,B_n}(o),$$

over a set $\mathcal{A}_n$ of $K_0(n)$ pairs (k,ϵ). The finite sample inequality of Corollary 1 holds for the cross-validated ϵ-net estimator $\hat{\Psi}(P_n) = \hat{\Psi}_{k(P_n),\epsilon(P_n)}(P_n)$, where k now plays the role of s.

Notice that a standard δ-grid for the coefficients β, corresponding to $\psi_\beta \in \mathbf{\Psi}_k$, consists of on the order of $\binom{N}{k}(1/\delta)^k$ points and that an ϵ-net

requires setting $\delta = \epsilon/k$. Thus, when a sequence of δ-grids is used to generate a sequence of ϵ-nets, we have

$$\begin{aligned}\log N_k(\epsilon) &= O\left(\log \frac{N(N-1)\ldots(N-k+1)}{k!} + k\log\frac{k}{\epsilon}\right) \\ &= O\left(\log \frac{N^k}{k!} + k\log\frac{k}{\epsilon}\right) \\ &= O(k\log N - k\log\epsilon).\end{aligned}$$

This yields the following finite sample inequality

$$\begin{aligned}&E\int(\hat{\Psi}_{k(P_n),\epsilon(P_n)}(P^0_{n,B_n})(w)-\psi_0(w))^2dF_0(w) \leq (1+2\lambda)\times \\ &\min_{(k,\epsilon)\in\mathcal{A}_n}\left\{(1+2\lambda)\min_{\psi\in\mathbf{\Psi}_{k,\epsilon}}\int(\psi(w)-\psi_0(w))^2dF_0(w) + \ 2C(\lambda)\tfrac{1+k\log(N/\epsilon)}{n(1-p)}\right\} \\ &\qquad\qquad + 2C(\lambda)\tfrac{1+\log(K_0(n))}{np}.\end{aligned}$$

If $\psi_0 \in \mathbf{\Psi}_{k^*}$ for some k^*, one may replace the minimum over k by simply k^* and note that the approximation error, $B_0(k^*,\epsilon) = \min_{\psi\in\mathbf{\Psi}_{k^*,\epsilon}}\int(\psi(w)-\psi_0(w))^2dF_0(w)$, of the ϵ-net $\mathbf{\Psi}_{k^*,\epsilon}$, is ϵ^2. Thus, the bound on the right-hand side becomes

$$O\left(\min_{\{\epsilon:(k^*,\epsilon)\in\mathcal{A}_n\}}\left\{\epsilon^2 + k^*\frac{\log(N(n))}{n} - k^*\frac{\log(\epsilon)}{n}\right\}\right) + O\left(\frac{\log(K_0(n))}{n}\right), \quad (37)$$

where we allow $N = N(n)$ and $k^* = k^*(n)$ to depend on the sample size n. Minimizing over ϵ yields

$$O\left(\frac{k^*}{n}\max\{\log(N(n)), \log(n/k^*(n))\}\right) + O\left(\frac{\log(K_0(n))}{n}\right).$$

Thus, up till a $\log(n)$ (or $\log(N(n))$) factor, we achieve the parametric rate of convergence corresponding with the space $\mathbf{\Psi}_{k^*}$. This demonstrates the adaptivity of the cross-validated ϵ-net estimator $\hat{\Psi}_{k(P_n),\epsilon(P_n)}(P_n)$.

Example 5 (Non-parametric regression: Adaptation to smoothness) Consider the regression setting as described in Corollary 1, where we assume that there exists a constant $C_0 < \infty$, so that $Pr(|Y| < C_0) = 1$. Let $\mathbf{\Psi} \equiv \{f : \| f \|_\infty < C_0\}$ be the non-parametric parameter space. Given constants M and α, let $\mathbf{\Psi}_{M,\alpha} \equiv \{f \in \mathbf{\Psi} : f \in C^\alpha_M\}$ be the subspaces of $\mathbf{\Psi}$ indexed by the smoothness degree α and a bound M on the α-derivative

(see equation (33)). For each (M, α, ϵ)-triple, let $\mathbf{\Psi}_{M,\alpha,\epsilon}$ be an ϵ-net of $\mathbf{\Psi}_{M,\alpha}$, consisting of $N_{M,\alpha}(\epsilon)$ points. Let $\hat{\Psi}_{M,\alpha,\epsilon}(P_n)$ be the *empirical* risk minimizer over the ϵ-net $\mathbf{\Psi}_{M,\alpha,\epsilon}$ and let $(M(P_n), \alpha(P_n), \epsilon(P_n))$ denote the minimizer of the *cross-validated* risk

$$(M, \alpha, \epsilon) \to E_{B_n} \int L(o, \hat{\Psi}_{M,\alpha,\epsilon}(P^0_{n,B_n}))dP^1_{n,B_n}(o),$$

over a set $\mathcal{A}_n$ of $K_0(n)$ triples (M, α, ϵ). The finite sample inequality of Corollary 1 holds for the cross-validated ϵ-net estimator $\hat{\Psi}(P_n) = \hat{\Psi}_{M(P_n),\alpha(P_n),\epsilon(P_n)}(P_n)$, where (M, α) now plays the role of s. Substitution for $N_{M,\alpha}(\epsilon)$ of the covering numbers of $\mathbf{\Psi}_{M,\alpha}$, as presented in Section 3.4, proves the asymptotic optimality and adaptivity of the proposed estimator $\hat{\Psi}(P_n)$.

4.2 Density estimation

Corollary 2 Density estimation.
Setting. *Let $O \sim P_0$, where $\psi_0 = f_0 \equiv \frac{dP_0}{d\mu}$ is the density of P_0 w.r.t. a dominating measure μ, and let $L(o, \psi) = -\log(\psi(o))$ denote the* negative log-density loss function.
Assumptions. *Assume that there exist constants $l > 0$ and $u < \infty$, so that for P_0-almost every O, $l < f_0(O) \leq u$ and $l <\mid \psi(O) \mid \leq u$, for all $\psi \in \mathbf{\Psi}$.*
Definitions. *Let $M_1 \equiv \log(u/l)$ and $M_2 \equiv 4u/l$.*
Finite sample result. *The finite sample inequality of Theorem 1 holds with constants M_1 and M_2 as defined above. That is, for any $\lambda > 0$,*

$$\begin{aligned} E \int -\log\left(\frac{\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n})(o)}{\psi_0(o)}\right) dP_0(o) \qquad (38) \\ \leq (1+2\lambda) \times \min_{(s,\epsilon)\in\mathcal{A}_n} \left\{ (1+2\lambda) \min_{\psi\in\mathbf{\Psi}_{s,\epsilon}} \int -\log\left(\frac{\psi(o)}{\psi_0(o)}\right) dP_0(o) \right. \\ \left. + \ 2C(\lambda)\frac{1+\log(N_s(\epsilon))}{n(1-p)} \right\} \\ + 2C(\lambda)\frac{1+\log(K_0(n))}{np}. \end{aligned}$$

Proof of Corollary 2. Application of Theorem 1 requires verification of Assumptions A1 and A2. Regarding Assumption A1, we note that

$$L(o, \psi) - L(o, \psi_0) = -\log\left(\frac{\psi(o)}{\psi_0(o)}\right).$$

Thus, Assumption A1 holds with $M_1 = \log(u/l)$. In Lemma 2, p. 9, van der Laan et al. (2004) show that Assumption A2 holds with $M_2 = 4u/l$. This proves the corollary.

□

We leave it to the reader to work out the analogues of the two regression examples above, for the estimation of a conditional density $f_0(Y \mid W)$. In this case, one constructs ϵ-nets corresponding with either different smoothness classes for the density $f_0(Y \mid W)$ or different degrees of sparsity for a particular high-dimensional parametric model (e.g., a Cox-proportional hazards model, linear in basis functions for W).

4.3 Multivariate outcome regression

Corollary 3 Multivariate outcome regression.
Setting. *Let $O = (W, Y) \sim P_0$, where $Y = (Y(l) : l = 1, \ldots, L)$ is a random outcome L-vector and W a vector of covariates with c.d.f. F_0. The parameter of interest is $\psi_0(W) \equiv E_{P_0}[Y \mid W] = (E_{P_0}[Y(l) \mid W] : l = 1, \ldots, L)$, the conditional expected value of the outcome vector Y given covariates W. Define the loss function as the* weighted quadratic loss function,

$$L(O, \psi) \equiv (Y - \psi(W))^\top \eta(W)(Y - \psi(W)), \tag{39}$$

where $\eta(\cdot)$ is a symmetric $L \times L$-matrix function of W, and note that, for any choice of $\eta(\cdot)$, the risk is minimized by the parameter value ψ_0, that is,

$$\psi_0 = argmin_{\psi \in \mathbf{\Psi}} E_{P_0} L(O, \psi) = argmin_{\psi \in \mathbf{\Psi}} \int L(o, \psi) dP_0(o). \tag{40}$$

The random (via the empirical distribution P_n) risk difference between the cross-validated ϵ-net estimator and the parameter value ψ_0 is given by

$$\begin{aligned} & E_{B_n} d_0(\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n}), \psi_0) \\ & \quad = E_{B_n} \int \left\| \eta^{1/2}(w) \left(\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n})(w) - \psi_0(w) \right) \right\|^2 dF_0(w), \end{aligned}$$

where $\| \cdot \|$ denotes the standard Euclidean norm, with $\| \vec{x} \|^2 \equiv \sum_{l=1}^L x_l^2$ for $\vec{x} \in \mathbb{R}^L$.

Assumptions. *Let* $C_0 \equiv \sup_{W,Y,l} \mid (\eta^{1/2}(W)Y)(l) \mid < \infty$. *For all* $\psi \in \mathbf{\Psi}$, *assume that* $\sup_{W,l} \mid (\eta^{1/2}(W)\psi(W))(l) \mid \leq C_0$. *If we define* $c(W) \equiv \sup_{\|\vec{x}\|=1} \parallel \eta^{1/2}(W)\vec{x} \parallel$ *as the matrix norm of the linear operator* $\eta^{1/2}(W) : \mathbb{R}^L \to \mathbb{R}^L$, *then we can choose* $C_0 \equiv \sup_W c(W) * \sup_{Y,l} \mid Y(l) \mid$. *Let* $C_1 < \infty$ *be so that for all* $\psi \in \mathbf{\Psi}$, *we have*

$$\sup_W \sum_{l=1}^{L} \mid (\eta^{1/2}(W)\psi(W))(l) \mid \leq C_1.$$

Definitions. *Let* $M_1 \equiv 5C_0C_1$ *and* $M_2 \equiv 16LC_0^2$. *If* W *is empty, then* $\eta(W) = \eta$ *is constant and we can set* $M_2 \equiv 4 \parallel \Sigma_0^{1/2} \parallel^2$, *where* $\Sigma_0 = COV_{P_0}[\eta^{1/2}Y]$ *and* $\parallel \Sigma_0^{1/2} \parallel = \sup_{\|\vec{x}\|=1} \parallel \Sigma_0^{1/2}\vec{x} \parallel$ *denotes the matrix norm of* $\Sigma_0^{1/2}$.

Finite sample result. *The finite sample inequality of Theorem 1 holds with constants* M_1 *and* M_2 *as defined above. That is, for any* $\lambda > 0$,

$$\begin{aligned} & E \int \left\| \eta^{1/2}(w) \left(\hat{\Psi}_{s(P_n),\epsilon(P_n)}(P^0_{n,B_n})(w) - \psi_0(w) \right) \right\|^2 dF_0(w) \qquad (41) \\ & \leq (1+2\lambda) \times \min_{(s,\epsilon)\in\mathcal{A}_n} \left\{ (1+2\lambda) \min_{\psi\in\mathbf{\Psi}_{s,\epsilon}} \int \left\| \eta^{1/2}(w) \left(\psi(w) - \psi_0(w) \right) \right\|^2 dF_0(w) \right. \\ & \qquad \left. + \ 2C(\lambda) \frac{1 + \log(N_s(\epsilon))}{n(1-p)} \right\} \\ & \quad + \ 2C(\lambda) \frac{1 + \log(K_0(n))}{np}. \end{aligned}$$

Proof of Corollary 3. We first derive the following general results concerning the weighted quadratic loss function. Observe that this loss function can be rewritten as

$$L(O,\psi) = (Y - \psi(W))^\top \eta(W)(Y - \psi(W)) = \parallel \eta^{1/2}(W)(Y - \psi(W)) \parallel^2,$$

and, for notational convenience, define the following three random L-vectors

$$\begin{aligned} \vec{a} &\equiv \eta^{1/2}(W)Y, \\ \vec{b} &\equiv \eta^{1/2}(W)\psi(W), \\ \vec{b}_0 &\equiv \eta^{1/2}(W)\psi_0(W). \end{aligned}$$

Then, the difference in loss functions at ψ and ψ_0 can be expressed as

$$\begin{aligned} L(O,\psi) - L(O,\psi_0) &= \| \eta^{1/2}(W)(Y - \psi(W)) \|^2 - \| \eta^{1/2}(W)(Y - \psi_0(W)) \|^2 \\ &= \sum_{l=1}^{L} (a_l - b_l)^2 - \sum_{l=1}^{L} (a_l - b_{0l})^2 \\ &= \sum_{l=1}^{L} (-2a_l(b_l - b_{0l}) + b_l^2 - b_{0l}^2), \end{aligned} \tag{42}$$

and the corresponding risk difference as

$$\begin{aligned} d_0(\psi,\psi_0) &= E_{P_0}\left[L(O,\psi) - L(O,\psi_0)\right] = \int (L(o,\psi) - L(o,\psi_0)) dP_0(o) \\ &= \int \| \eta^{1/2}(w)(\psi(w) - \psi_0(w)) \|^2 \, dF_0(w) = E_{P_0} \| \vec{b} - \vec{b}_0 \|^2 \end{aligned} \tag{43}$$

Regarding Assumption A1, recall that $\max_l \mid a_l \mid \leq C_0$, $\max_l \mid b_l \mid \leq C_0$, $\sum_l \mid b_l \mid \leq C_1$, and $\sum_l \mid b_{0l} \mid \leq C_1$. It then follows, from the last expression in equation (42), that $L(O,\psi) - L(O,\psi_0)$ is bounded by $5C_0C_1$. Hence, Assumption A1 indeed holds with $M_1 = 5C_0C_1$.

Regarding Assumption A2, first consider the case that W is empty. In the representation of $L(O,\psi) - L(O,\psi_0)$ as $\sum_l (-2a_l(b_l - b_{0l}) + b_l^2 - b_{0l}^2)$ in equation (42), we note that only $a_l = (\eta^{1/2}Y)(l)$ is random. Thus,

$$\begin{aligned} \mathrm{VAR}_{P_0}\left[L(O,\psi) - L(O,\psi_0)\right] &= \mathrm{VAR}_{P_0}\left[\sum_{l=1}^{L} 2(b_l - b_{0l})(\eta^{1/2}Y)(l)\right] \\ &= 4(\vec{b} - \vec{b}_0)^\top (\mathrm{COV}_{P_0}[\eta^{1/2}Y])(\vec{b} - \vec{b}_0) \\ &= 4(\vec{b} - \vec{b}_0)^\top \Sigma_0 (\vec{b} - \vec{b}_0) \\ &= 4 \| \Sigma_0^{1/2}(\vec{b} - \vec{b}_0) \|^2 \\ &\leq 4 \| \Sigma_0^{1/2} \|^2 \| \vec{b} - \vec{b}_0 \|^2 \\ &= M_2 \, E_{P_0}\left[L(O,\psi) - L(O,\psi_0)\right], \end{aligned}$$

which proves Assumption A2 with $M_2 = 4 \| \Sigma_0^{1/2} \|^2$. Let us now consider the general case, where W is a random vector of covariates. Define $e(O) \equiv \eta^{1/2}(W)(Y - \psi(W))$ and $e_0(O) \equiv \eta^{1/2}(W)(Y - \psi_0(W))$. Then, we have the

following representation for the difference in loss functions at ψ and ψ_0

$$L(O,\psi)-L(O,\psi_0)=\sum_{l=1}^{L}(e_l^2(O)-e_{0l}^2(O))=\sum_{l=1}^{L}(e_l(O)-e_{0l}(O))(e_l(O)+e_{0l}(O)),$$

and hence

$$\begin{aligned}
E_{P_0}\left[(L(O,\psi)-L(O,\psi_0))^2\right] &= \int (L(o,\psi)-L(o,\psi_0))^2\, dP_0(o)\\
&= \int \left(\sum_{l=1}^{L}(e_l(o)-e_{0l}(o))(e_l(o)+e_{0l}(o))\right)^2 dP_0(o)\\
&\leq \int \sum_{l=1}^{L}(e_l(o)-e_{0l}(o))^2 \sum_{l=1}^{L}(e_l(o)+e_{0l}(o))^2\, dP_0(o)\\
&\leq \sup_o \parallel e(o)+e_0(o) \parallel^2 \int \sum_{l=1}^{L}(e_l(o)-e_{0l}(o))^2\, dP_0(o)\\
&= \sup_o \parallel e(o)+e_0(o) \parallel^2 \int \parallel \eta^{1/2}(\psi(w)-\psi_0(w)) \parallel^2\, dF_0(w)\\
&= \sup_o \parallel e(o)+e_0(o) \parallel^2\ E_{P_0}[L(O,\psi)-L(O,\psi_0)]\\
&\leq 16LC_0^2\ E_{P_0}[L(O,\psi)-L(O,\psi_0)].
\end{aligned}$$

The last equality follows from equation (43) and the last inequality from two applications of $\parallel a+b \parallel^2 \leq 4\max(\parallel a \parallel^2, \parallel b \parallel^2)$. This proves that Assumption A2 holds with $M_2 = 16LC_0^2$. We have now verified both Assumptions A1 and A2. This completes the proof of the corollary.

□

Example 6 (Estimation of a multivariate mean: Adaptation to sparsity) Consider the random L-vector $O = Y = (Y(l) : l = 1,\ldots,L) \sim P_0$ and let the mean L-vector $\psi_0 \equiv E_{P_0}[Y] = (E_{P_0}[Y(l)] : l = 1,\ldots,L)$ denote the parameter we wish to estimate. For a candidate L-vector ψ, define as loss function

$$L(O,\psi) \equiv \sum_{l=1}^{L}\eta(l)(Y(l)-\psi(l))^2,$$

where $\eta(l) = 1/\sigma^2(l)$, $l = 1, \ldots, L$, is a given set of weights. For some $C_1 < \infty$, let

$$\mathbf{\Psi} \equiv \left\{ \psi \in \mathbb{R}^L : \sum_{l=1}^{L} \mid \psi(l)/\sigma(l) \mid \leq C_1 \right\}$$

be the assumed parameter space. Let $\mathbf{\Psi}_k \equiv \{\psi \in \mathbf{\Psi} : \sum_{l=1}^{L} \mathrm{I}(\psi(l) \neq 0) \leq k\}$ consist of all L-vectors in $\mathbf{\Psi}$ with at most k non-zero components, $k = 1, \ldots, L$. For each (k, ϵ)-pair, let $\mathbf{\Psi}_{k,\epsilon}$ be an ϵ-net of $\mathbf{\Psi}_k$, consisting of $N_k(\epsilon)$ points. Let $\hat{\Psi}_{k,\epsilon}(P_n)$ be the *empirical* risk minimizer over the ϵ-net $\mathbf{\Psi}_{k,\epsilon}$ and let $(k(P_n), \epsilon(P_n))$ denote the minimizer of the *cross-validated* risk

$$\begin{aligned} (k, \epsilon) &\rightarrow E_{B_n} \int L(o, \hat{\Psi}_{k,\epsilon}(P^0_{n,B_n}))dP^1_{n,B_n}(o) \\ &= \sum_{l=1}^{L} E_{B_n} \frac{1}{n(1-p)} \sum_{\{i:B_n(i)=1\}} \eta(l)(Y_i(l) - \hat{\Psi}_{k,\epsilon}(P^0_{n,B_n})(l))^2, \end{aligned}$$

over a set $\mathcal{A}_n$ of $K_0(n)$ pairs (k, ϵ). The finite sample inequality of Corollary 3 holds for the cross-validated ϵ-net estimator $\hat{\Psi}(P_n) = \hat{\Psi}_{k(P_n),\epsilon(P_n)}(P_n)$, where k now plays the role of s, $C_0 \equiv \sup_{Y,l} \mid Y(l)/\sigma(l) \mid < \infty$, $M_1 = 5C_0C_1$, and $M_2 = 4 \parallel \Sigma_0^{1/2} \parallel^2$. Here, $\Sigma_0 = \mathrm{COV}_{P_0}[Y/\sigma]$ is the covariance matrix of $(Y - \psi_0)/\sigma$ and $\parallel \Sigma_0^{1/2} \parallel$ denotes the matrix norm of its square root.

Notice, as in Example 4 for univariate outcome regression, that a δ-grid for $\mathbf{\Psi}_k$ consists of on the order of $\binom{L}{k}(1/\delta)^k$ points and that an ϵ-net requires setting $\delta = \epsilon/k$. Thus, when a sequence of δ-grids is used to generate a sequence of ϵ-nets, we have

$$\log N_k(\epsilon) = O(k \log L - k \log \epsilon).$$

This yields the following finite sample inequality

$$\begin{aligned} &E \sum_{l=1}^{L} (\hat{\Psi}_{k(P_n),\epsilon(P_n)}(P^0_{n,B_n})(l) - \psi_0(l))^2 \tfrac{1}{\sigma^2(l)} \leq (1 + 2\lambda) \times \\ &\min_{(k,\epsilon)\in\mathcal{A}_n} \left\{ (1 + 2\lambda) \min_{\psi\in\mathbf{\Psi}_{k,\epsilon}} \sum_{l=1}^{L} (\psi(l) - \psi_0(l))^2 \tfrac{1}{\sigma^2(l)} + 2C(\lambda) \tfrac{1 + k \log(L/\epsilon)}{n(1-p)} \right\} \\ &\qquad + 2C(\lambda) \tfrac{1 + \log(K_0(n))}{np}. \end{aligned}$$

If $\psi_0 \in \mathbf{\Psi}_{k^*}$ for some k^*, then straightforward algebra, as on p. 29 for Example 4, shows that the upper bound on the right-hand side is bounded by

$$O\left(\frac{k^*}{n} \max\{\log(L(n)), \log(n/k^*(n))\}\right) + O\left(\frac{\log(K_0(n))}{n}\right),$$

where we allow $L = L(n)$ and $k^* = k^*(n)$ to depend on the sample size n. Thus, up till a $\log(n)$ (or $\log(L(n))$) factor, we achieve the parametric rate of convergence corresponding with the space $\mathbf{\Psi}_{k^*}$. This demonstrates the adaptivity of the cross-validated ϵ-net estimator $\hat{\Psi}_{k(P_n),\epsilon(P_n)}(P_n)$.

5 Simulation study

Datasets were simulated as i.i.d. realizations from a linear regression model, $Y \sim \beta W + \epsilon$, where $W = (W(j) : j = 1, \ldots, 10)$ is a 10-dimensional vector of covariates, with $W(j)$ i.i.d. $U(0, 1)$, the error term $\epsilon \sim N(0, \sigma^2)$, and ϵ and W are independent. The regression coefficients β_j, $j = 1, \ldots, 10$, were generated as i.i.d. $U(-5, 5)$ random variables. Ten datasets of various sample sizes, $n = 20, 30, 40, 50, 70, 90, 120$, were simulated from this model. We compared the performance (in terms of conditional risk) of the cross-validated ϵ-net linear regression estimator of β, based on a standard δ-grid as described in Section 2.2.2, to the performance of the least angle linear regression estimator (Least Angle Regression, or LARS, in Efron et al. (2004)) and L^1-penalized least squares linear regression estimator (Least Absolute Shrinkage and Selection Operator, or LASSO, in Hastie et al. (2001) and Tibshirani (1996)). The latter two estimators were obtained using the default version of the `lars` function from the R package `lars` (Hastie and Efron, 2003). The `lars` function selects the penalty term in LASSO with cross-validation. The three estimators were evaluated based on their mean squared residual errors on an independent test set of 10,000 observations, which approximates the true conditional risk for the quadratic loss function. The results are reported in Tables 1 and 2, corresponding to choices $\sigma^2 = 1$ and $\sigma^2 = 100$, respectively, for the variance of the error term ϵ. Note that the error variance σ^2 equals the optimal risk achieved by the true regression function, $\psi_0(W) = \beta_0 W$. One therefore seeks estimators whose risk is as close as possible to σ^2. These preliminary simulation results show that the proposed cross-validated ϵ-net estimator outperforms both LARS and LASSO in this simple linear regression setting with independent covariates.

We are planning more extensive simulation studies to further investigate the properties and practical performance of the cross-validated adaptive ϵ-nest estimator. We note that in the case of correlated covariates, one could construct δ-grids for orthonormalized covariate vectors, as discussed in Section 2.2.1. Since orthogonalizing covariates is generally considered a good

strategy for prediction purposes, we suspect that the LASSO would certainly not be harmed by such an orthonormalization step. Thus, we expect the results in Tables 1 and 2 to also hold for the ϵ-net and LASSO estimators applied to orthonormalized covariate vectors.

6 Discussion

We have proposed a cross-validated adaptive ϵ-net estimation methodology that covers a broad class of estimation problems, including multivariate outcome prediction and multivariate density estimation. In this approach, one considers collections of ϵ-net sieves, corresponding to different parameterizations, or subspaces $\mathbf{\Psi}_s$, of the parameter space $\mathbf{\Psi}$, where the ϵ-nets $\mathbf{\Psi}_{s,\epsilon}$ provide arbitrary good approximations of the parameter subspaces $\mathbf{\Psi}_s$. For each choice of subspace $\mathbf{\Psi}_s$ and resolution ϵ, one generates candidate estimators as the empirical risk minimizers over the ϵ-nets $\mathbf{\Psi}_{s,\epsilon}$. The proposed cross-validated ϵ-net estimator is the candidate estimator corresponding to the choice of subspace and ϵ-value minimizing the cross-validated risk. The finite sample inequalities of Theorems 1 and 3 prove that the cross-validated ϵ-net estimator achieves at worst the minimax rate of convergence and is highly adaptive to the true parameter value ψ_0.

Current practice typically differs from our proposed approach in four important ways: 1) one usually selects continuous (infinite) sieves; 2) the elements of the sieves are typically nested; 3) one often adheres to forward/backward-type search algorithms, which do not even attempt to minimize the empirical risk over a given parameter subspace; and 4) one typically considers only one sieve, indexed by one complexity parameter (usually the size of the subspace), instead of using cross-validation to select among a class of sieves for subspaces and/or the complete parameter space (e.g., choice of basis). We plan to carry out a simulation study to investigate the effects of these four issues in the context of regression.

Regarding 4), our finite sample results show that choosing a rich collection of ϵ-net sieves (e.g., containing arbitrary good approximations of guessed subspaces) will typically only improve the estimator's finite sample and asymptotic performance. In particular, if the ϵ-nets are indexed by a choice of basis for the parameter space (or subspace), then our proposed estimator data-adaptively chooses a basis which is most effective in approximating the true parameter value ψ_0. Specifically, general Theorems 2 and 4

demonstrate that cross-validation is as good (up till a typically second order term, $\log(K(n))/n$, for the expected risk difference) in selecting among a collection of $K(n)$ estimators as an oracle procedure which makes an optimal P_0-dependent choice *for the given dataset*. This shows that in large models, the number $K(n)$ of candidate estimators, or ϵ-nets, can typically be very large, e.g., $K(n) = n^m$ for some $m < \infty$.

Issue 3), namely, the observation that many of the algorithms used in regression are not aiming to minimize empirical risk over specified parameter spaces (e.g., forward variable selection, recursive partitioning in classification and regression trees), has motivated us to develop more aggressive *deletion/substitution/addition* or *D/S/A algorithms* (Molinaro and van der Laan, 2004; Sinisi and van der Laan, 2004). These D/S/A algorithms truly aim to minimize risk over all regression functions with maximally k basis functions, while one still selects k and other fine-tuning parameters (such as the complexity of basis functions and constraints on the coefficients) with cross-validation. Initial simulation results show that these aggressive algorithms are more adaptive and can easily outperform forward selection and recursive partitioning algorithms.

van der Laan and Dudoit (2003, 2004) provide an important generalization of this loss-based estimation framework, by allowing the loss function to depend on a nuisance parameter υ (i.e., $L(O, \psi \mid \upsilon)$) and extending the above estimation procedure and theorems to this case. This generalization now covers regression with censored data, density estimation with censored data, causal inference, and many other applications. In censored data situations, the loss function can be chosen to be the (double robust) inverse probability of censoring weighted full data loss function, as presented in van der Laan and Robins (2002).

Acknowledgement. We thank Peter Dimitrov for carrying out the simulation study presented in this article.

Table 1: *Risk for cross-validated ϵ-net estimator, LASSO, and LARS ($\sigma^2 = 1$).* For each of the three estimators, we report the average and standard deviation (over ten simulations) of the test set risk (approximating the true conditional risk) for the squared error loss function. The test set consists of 10,000 i.i.d. observations simulated from the linear regression model, $Y \sim \beta W + \epsilon$, with ten i.i.d. $U(0,1)$ covariates W and independent errors $\epsilon \sim N(0, \sigma^2 = 1)$. Five-fold cross-validation was used for all estimators.

n	**ϵ-net**		**LARS**		**LASSO**	
	mean.eps.net	sd.eps.net	mean.lars	sd.lars	mean.lasso	sd.lasso
20	2.11	0.33	261.75	225.62	111.70	182.26
30	1.53	0.22	2.02	0.83	3.80	2.91
40	1.70	0.44	1.49	0.20	1.50	0.22
50	1.21	0.06	1.76	0.44	1.96	0.35
70	1.18	0.04	2.96	1.51	2.96	1.51
90	1.17	0.10	1.48	0.09	1.49	0.07
120	1.04	0.01	1.12	0.09	1.12	0.09

Table 2: *Risk for cross-validated ϵ-net estimator, LASSO, and LARS ($\sigma^2 = 100$).* For each of the three estimators, we report the average and standard deviation (over ten simulations) of the test set risk (approximating the true conditional risk) for the squared error loss function. The test set consists of 10,000 i.i.d. observations simulated from the linear regression model, $Y \sim \beta W + \epsilon$, with ten i.i.d. $U(0,1)$ covariates W and independent errors $\epsilon \sim N(0, \sigma^2 = 100)$. Five-fold cross-validation was used for all estimators.

n	**ϵ-net**		**LARS**		**LASSO**	
	mean.eps.net	sd.eps.net	mean.lars	sd.lars	mean.lasso	sd.lasso
20	196.25	28.69	521.15	68.06	603.04	208.31
30	142.14	20.48	229.40	134.10	222.86	102.17
40	135.03	12.37	186.13	49.65	166.42	19.48
50	111.96	1.38	216.72	16.98	215.66	37.49
70	115.69	5.53	372.41	246.64	346.72	189.70
90	116.22	3.85	138.61	12.54	149.31	10.22
120	110.33	6.01	106.41	1.59	109.00	3.87

References

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

A. R. Barron. Statistical properties of artificial neural networks. In *Proceedings of the 28th conference on decision theory and control, Tampa, Florida*, December 1989.

A. R. Barron. *Nonparametric functional estimation and related topics*, chapter Complexity regularization with application to artificial neural networks, pages 561–576. Kluwer Academic Publishers, the Netherlands, 1991.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE transactions on information theory*, 39(3):930–945, 1993.

L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York, 1997.

L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984.

C. Canuto and A. Quarteroni. Approximation results for orthogonal polynomials in Sobolev spaces. *Mathematics of Computation*, 38(157):67–86, 1982.

D. L. Donoho. Le Cam Lecture: Estimation by ϵ-nets. *IMS Le Cam Lecture 2003 at the Joint Statistical Meeting, San Francisco*, 2003.

D. L. Donoho and I. M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of basis. *C.R. Acad. Sci. Paris, Ser I*, 319:1317–1322, 1994.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in model selection and performance assessmen. Technical Report 126, Division of Biostatistics, University of California, Berkeley, 2003. URL `www.bepress.com/ucbbiostat/paper126/`.

S. Dudoit, M. J. van der Laan, S. Keleş, A. M. Molinaro, S. E. Sinisi, and S. L. Teng. Loss-based estimation with cross-validation: Applications to microarray data analysis. *SIGKDD Explorations, Microarray Data Mining Special Issue*, 2004. (To appear).

B. Efron, T. Hastie, I. Johnstone, and R. Tibhirani. Least angle regression. *Annals of Statistics*, 32(4), 2004. (To appear).

T. Hastie and B. Efron. Least angle regression, lasso and forward stagewise. R package, 2003. URL `cran.r-project.org`. Version 0.9-4.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

I. M. Johnstone. Oracle inequalities and nonparametric function estimation. *Journal der Deutchen Mathematiker Vereinigung, Proc. of the International Congress of Mathematicians, Berlin 1998*, III:267–278, 1998.

S. Keleş, M. J. van der Laan, and S. Dudoit. Asymptotically optimal model selection method for regression on censored outcomes. Technical Report 124, Division of Biostatistics, University of California, Berkeley, 2003. URL `www.bepress.com/ucbbiostat/paper124/`.

L. M. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, 1986.

L. M. Le Cam and G. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag, New York, 1990.

M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1996.

P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. Technical report, Department of Mathematics, Paris-Sud, 1998.

A. M. Molinaro, S. Dudoit, and M. J. van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 2004. (To appear).

A. M. Molinaro and M. J. van der Laan. A Deletion/Substitution/Addition algorithm for partitioning the covariate space in prediction. Technical report, Division of Biostatistics, UC Berkeley, 2004. (In preparation).

B. D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, New York, 1996.

S. Sinisi and M. J. van der Laan. A general Deletion/Substitution/Addition algorithm in prediction. Technical report, Division of Biostatistics, UC Berkeley, 2004. (In preparation).

M. Talagrand. A new look at independence. *Ann. Probab.*, 24:1–34, 1996a.

M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563, 1996b.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive ϵ-net estimator: Finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, 2003. URL `www.bepress.com/ucbbiostat/paper130/`.

M. J. van der Laan and S. Dudoit. *Loss-Based Estimation*. 2004. (In preparation).

M. J. van der Laan, S. Dudoit, and S. Keleş. Asymptotic optimality of likelihood based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 2004. (To appear).

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York, 2002.

A. W. van der Vaart. A note on cross-validation theory articles by Sandrine Dudoit, Mark van der Laan, and Maarten Wegkamp. Technical report, Department of Statistics, Free University Amsterdam, 2003.

A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

Y. Yang and A. R. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.